

United States  
Department of  
Agriculture

Forest Service

Rocky Mountain  
Forest and Range  
Experiment Station

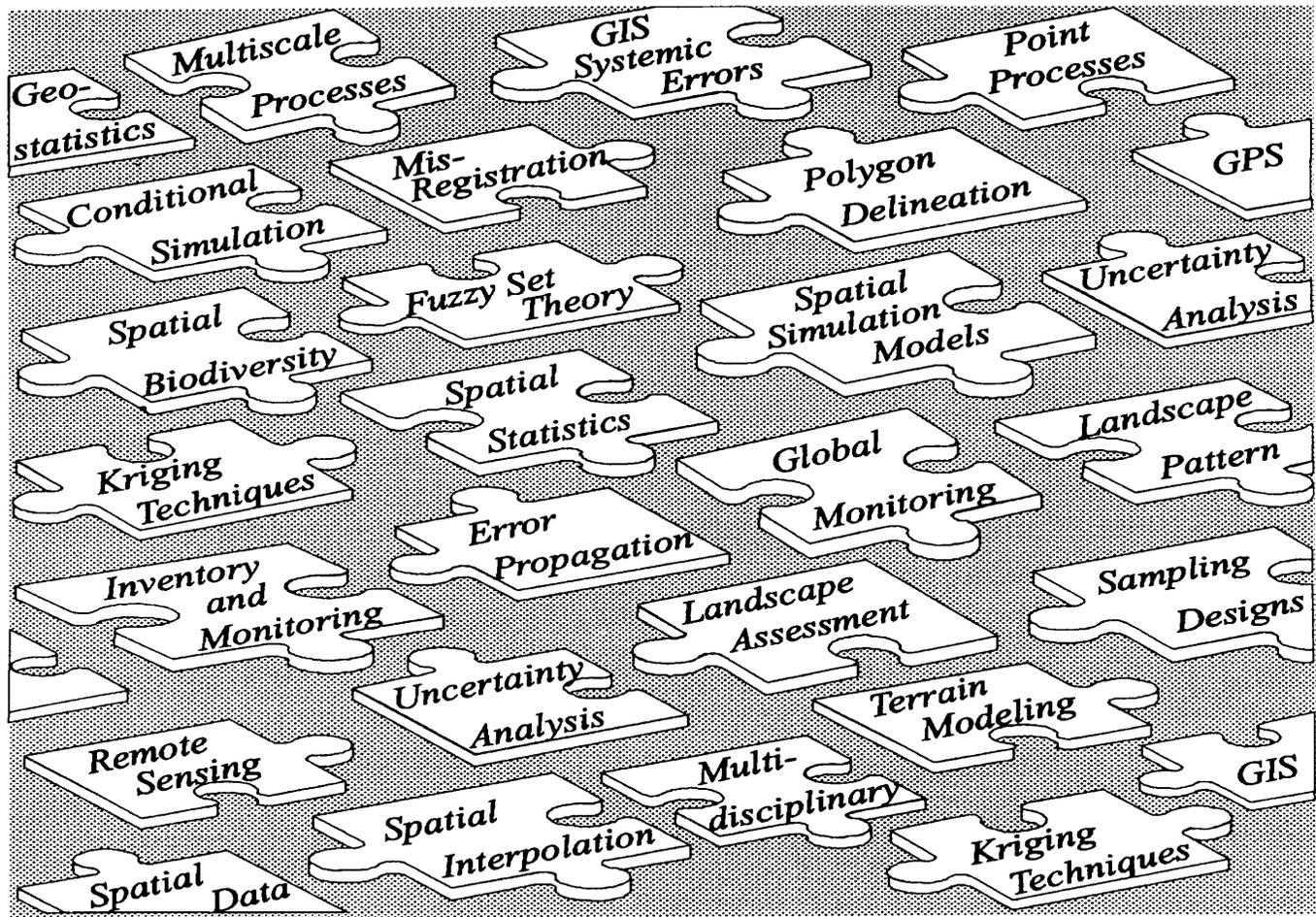
Fort Collins,  
Colorado 80526

General Technical  
Report RM-GTR-277

# Spatial Accuracy Assessment in Natural Resources and Environmental Sciences: Second International Symposium

May 21-23, 1996  
Fort Collins, Colorado

H. Todd Mowrer, Raymond L. Czaplewski, R. H. Hamre



## Abstract

Mowrer, H. Todd; Czaplewski, Raymond L.; Hamre, R. H. tech coords. 1996. Spatial Accuracy Assessment in Natural Resources and Environmental Sciences: Second International Symposium. May 21-23, 1996. General Technical Report RM-GTR-277. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Forest and Range Experiment Station. 728 p.

This international symposium on theory and techniques for assessing the accuracy of spatial data and spatial analyses included more than ninety presentations by representatives from government, academic, and private institutions in over twenty countries throughout the world. To encourage interactions across disciplines, presentations in the general subject areas of spatial statistics, geographic information systems, remote sensing, and multidisciplinary approaches were intermixed throughout the three days of sessions.

## Primary Sponsor

Rocky Mountain Forest and Range Experiment Station  
United States Department of Agriculture, Forest Service  
Fort Collins, Colorado

## Endorsing Organizations

International Union of Forest Research Organizations (IUFRO):  
Sections 4.01, 4.02, 4.11, and 4.12  
American Statistical Association: Section on Statistics and the Environment  
National Center for Geographic Information Analysis  
International Biometric Society  
American Society for Photogrammetry and Remote Sensing: GIS Division,  
European Association of Remote Sensing Laboratories:  
Special Interest Group on Landuse Planning and Forestry  
United States Environmental Protection Agency

### Copies of this publication may be obtained by the following methods:

**E-mail:** /s=r.schneider/oul=s28a@mhs-fswa.attmail.com  
**Phone:** (970) 498-1719  
**FAX:** (970) 498-1660 Attn: Publications  
**Mail:** Publications Distribution  
Rocky Mountain Forest and Range Experiment Station  
3825 East Mulberry Street  
Fort Collins, CO 80524-8597

**Spatial Accuracy Assessment in  
Natural Resources and Environmental Sciences:  
Second International Symposium**

**Technical Coordinators:**

**H. Todd Mowrer, Raymond L. Czaplewski, R. H. Hamre  
Rocky Mountain Forest and Range Experiment Station<sup>1</sup>**

<sup>1</sup>Headquarters is in Fort Collins, in cooperation with Colorado State University.

# Preface

This Second International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences follows two years after the first symposium in Williamsburg, Virginia, USA, chaired by Dr. James L. Smith. We hope to maintain and build on the spirit of open discussion and considerate but honest evaluation that Jim engendered throughout the Williamsburg Symposium.

It is apparent from the size of these proceedings that spatial accuracy assessment is an area of increasing effort and importance. Within the past decade, we have developed extensive spatial databases and have begun to realize their potential. We have now reached a point of maturity in spatial analyses where we must *routinely* evaluate the quality of those data sets, and the consequent reliability of the end-products of their use. Spatial data, and model predictions of spatial attributes, are becoming widely recognized as necessary support for intelligent decisions in resource management and environmental protection. Decision-makers and their staffs are becoming more accountable for evaluation of the quality of spatial data and model predictions, upon which policy analyses, strategic plans, and risk assessments are built. Objective evaluations of spatial accuracy are necessary to compare alternative technologies and competing models. Such comparisons speed improvement in our knowledge and in the state-of-the-art.

In order to further the informed use of spatial data and to improve data quality, the natural resource and environmental science communities must use a multidisciplinary approach, drawing from the methodologies presented here. We have attempted to build upon the theme of the first symposium, "unlocking the puzzle," through the logo of puzzle pieces used on the cover of these proceedings, indicating the many techniques and subject areas that must be intermeshed to produce a coherent picture of spatial accuracy assessment.

The Steering Committee wishes to express our deep appreciation to Robert H. Hamre, retired Station Editor, who kindly volunteered his services to professionally format these proceedings. Bob has generously and cheerfully given the proceedings a level of quality that otherwise would not have been within our resources. The papers included in these proceedings were produced in camera-ready format by the individual authors, who have sole responsibility for their content. By virtue of their effort, participation, and support, it is our hope that these proceedings may provide a significant contribution to further the recognition and development of spatial accuracy assessment in natural resources and environmental sciences.

## Contents

	Page
Spatial Statistics	
<b>Statistical modeling of environmental data in space and time .....</b>	<b>1</b>
<i>Noel Cressie</i>	
<b>A case study in geostatistical modeling for petroleum reservoir description .....</b>	<b>4</b>
<i>Jeffrey M. Yarus, Jeffrey A. May, and Timothy C. Coburn</i>	
<b>An overview of stochastic spatial simulation .....</b>	<b>13</b>
<i>R. Mohan Srivastava</i>	
<b>Choosing and using simulation algorithms .....</b>	<b>23</b>
<i>Donald E. Myers</i>	
<b>The components of geostatistical simulation .....</b>	<b>30</b>
<i>Carol A. Gotway and Brian M. Rutherford</i>	
<b>A general method for the efficient selection of sampling locations for problems in environmental restoration .....</b>	<b>39</b>
<i>Brian M. Rutherford</i>	
<b>Consideration of spatial variability in the management of non-point source pollution to groundwater .....</b>	<b>49</b>
<i>W. Woldt, F. Goderya, M. Dahab, and I. Bogardi</i>	
<b>Incorporating soil variability into a spatially distributed model of percolate accounting .....</b>	<b>57</b>
<i>Andrew S. Rogowski</i>	
<b>Response of a watershed model to varying spatial landscape characteristics .....</b>	<b>65</b>
<i>Dennis P. Swaney, Wen Ling Kuo, David A. Weinstein, Charles Mohler, Stephen DeGloria, Chris Pelkie, Fuan Tsai, Tammo S. Steenhuis, and Charles E. McCulloch</i>	
<b>Understanding the spatial distribution of tree species in Pennsylvania .....</b>	<b>73</b>
<i>Rachel Riemann Hershey</i>	
<b>A calibration-based model for correcting area estimates from coarse resolution land cover data .....</b>	<b>83</b>
<i>Aaron Moody</i>	
<b>Geostatistical analysis of multi-spatial resolution imaging spectrometer data for characterizing forest ecosystems .....</b>	<b>91</b>
<i>Paul M. Treitz and Philip J. Howarth</i>	
<b>Application of multiscale modeling for characterizing variability in hydrologic processes .....</b>	<b>99</b>
<i>Praveen Kumar</i>	
<b>Flow based scale-up of heterogeneous porous media using homogenization and wavelet representation .....</b>	<b>107</b>
<i>Joe Koebbe and Ryan Thomas</i>	
<b>Nonhomogeneous hidden Markov models allowing stochastic downscaling of synoptic atmospheric patterns to local hydrologic phenomena .....</b>	<b>115</b>
<i>Peter Guttorp and James P. Hughes</i>	
<b>Explicit consideration of multiple landscape scales while selecting spatial resolutions .....</b>	<b>121</b>
<i>John B. Collins and Curtis E. Woodcock</i>	

*(Continued)*

<b>Using trend surface methodology to compare spatial surfaces .....</b>	<b>129</b>
<i>Dale Zimmerman, Zhi-Jun Liu, and George Hallberg</i>	
<b>The effect of spatial covariance heterogeneity on prediction variance .....</b>	<b>137</b>
<i>J. Andrew Royle and Doug Nychka</i>	
<b>Further explorations of relationships between semi-variogram and spatial autoregressive models .....</b>	<b>147</b>
<i>Daniel A. Griffith, Larry J. Lane, and Philip G. Doyle</i>	
<b>A gamma-function model for d-variate spectra and cross-spectra for large scale frequency domain simulation of stationary random functions in <math>R^n</math> .....</b>	<b>155</b>
<i>Leon F. Borgman and John W. Kern</i>	
<b>Computing the area affected by phosphorus runoff in an Everglades wetland using Bayesian and universal kriging .....</b>	<b>164</b>
<i>Song S. Qian</i>	
<b>Mapping synecological coordinates: a spatial analysis of environmental indices in a forested landscape .....</b>	<b>173</b>
<i>Margaret R. Holdaway and Gary J. Brand</i>	
<b>On the robustness of the data assimilation methods in air pollution modeling .....</b>	<b>181</b>
<i>X. F. Zhang and A. W. Heemink</i>	
<b>Geographic Information Systems</b>	
<b>The development of a GIS information quality module .....</b>	<b>189</b>
<i>Jane Drummond, Allan Brown, Du Daosheng, Corné van Elzakker</i>	
<b>Communication of uncertainty in spatial data to policy makers .....</b>	<b>199</b>
<i>Morwenna Spear, Jane Hall, and Richard Wadsworth</i>	
<b>Measuring the performance of algorithms for generating ground slope .....</b>	<b>208</b>
<i>William H. Ryder and Demetra E. Voyadgis</i>	
<b>Experimental development of a model of vector data uncertainty .....</b>	<b>217</b>
<i>G. J. Hunter, J. Höck, M. Robey, and M. F. Goodchild</i>	
<b>A user-friendly tool for error modelling and error propagation in a GIS environment .....</b>	<b>225</b>
<i>Frank Forier and Frank Canters</i>	
<b>Discrete polygons or a continuous surface: which is the appropriate way to model forests cartographically? .....</b>	<b>235</b>
<i>Kim Lowell</i>	
<b>Choosing between abrupt and gradual spatial variation? .....</b>	<b>243</b>
<i>Gerard B. M. Heuvelink and Johan A. Huisman</i>	
<b>Stratified kriging using vague transition zones .....</b>	<b>251</b>
<i>G. Boucneau, M. Van Meirvenne, O. Thas, and G. Hofman</i>	
<b>The effect of uncertain locations on disease cluster statistics .....</b>	<b>259</b>
<i>Geoffrey M. Jacquez and Lance A. Waller</i>	
<b>Reliability of area mapping by delineation in aerial photographs .....</b>	<b>267</b>
<i>Claus-Peter Gross and Petra Adler</i>	
<b>The effect of database generalization on the accuracy of the viewshed .....</b>	<b>272</b>
<i>Peter Fisher</i>	
<b>Modeling slope stability uncertainty: a case study at the H. J. Andrews Long-Term Ecological Research Site, Oregon .....</b>	<b>281</b>
<i>Michelle L. Murillo and Gary J. Hunter</i>	

(Continued)

<b>An accuracy assessment of a spatial bioclimatic model .....</b>	<b>291</b>
<i>D. W. McKenney, B. G. Mackey, M. F. Hutchinson, and R. A. Sims</i>	
<b>An assessment of the horizontal accuracy of interim terrain data .....</b>	<b>301</b>
<i>L. A. Fatale, J. A. Messmore, and J. R. Ackeret</i>	
<b>Acquisition of spatial data by forest management agencies: a review .....</b>	<b>309</b>
<i>Michael J. C. Weir</i>	
<b>Attribute and positional accuracy assessment of the Murray Darling Basin project, Australia .....</b>	<b>317</b>
<i>R. W. Fitzgerald, K. T. Ritman, and A. Lewis</i>	
<b>A new method for evaluating positional map accuracy .....</b>	<b>327</b>
<i>Michael Lodin and David Skea</i>	
<b>Exploring spatial confidence with a raster-based GIS .....</b>	<b>336</b>
<i>Matthew H. Pelkki</i>	
<b>GPS-located accuracy assessment plots on the Modoc National Forest .....</b>	<b>343</b>
<i>Kevin Casey</i>	
<b>A method for measuring the spatial accuracy of coordinates collected using the global positioning system .....</b>	<b>351</b>
<i>Thomas Owens and David McConville</i>	
<b>GPS vs. traditional methods of data input: improving spatial data accuracy? .....</b>	<b>359</b>
<i>Russell Combs, Jr., James L. Smith, and Paul V. Bolstad</i>	

#### Remote Sensing

<b>Large scale tropical forest change monitoring using multiple resolution satellite data: from hot spot detection to global deforestation assessment? .....</b>	<b>366</b>
<i>Hervé Jeanjean, Frédéric Achard, and Jean-Paul Malingreau</i>	
<b>Optimum area sampling frame using high resolution satellite images with operational objective: how to conciliate statistical requirements and practical aspects .....</b>	<b>375</b>
<i>Hélène De Boissezon and Hervé Jeanjean</i>	
<b>The use of AVHRR satellite imagery to monitor boreal ecosystem forest fires .....</b>	<b>383</b>
<i>Donald R. Cahoon, Jr., and Brian J. Stocks</i>	
<b>Large area forest cover assessment: effects of misregistration in a double sampling approach with coarse and high resolution satellite images .....</b>	<b>391</b>
<i>Christoph Kleinn, Berthold Traub, and Matthias Dees</i>	
<b>Linear mixture modeling with autocorrelated errors .....</b>	<b>401</b>
<i>Jayantha Ediriwickrema, Siamak Khorram, Marcia Gumpertz, and John Brockhaus</i>	
<b>Validation process of the USGS seasonal land cover regions of the conterminous United States .....</b>	<b>409</b>
<i>Zhiliang Zhu, Donald O. Ohlen, Raymond L. Czaplewski, and Robert E. Burgan</i>	
<b>Investigation of possible contributions NDVI's have to misclassification in AVHRR data analysis .....</b>	<b>419</b>
<i>David L. Evans and Raymond L. Czaplewski</i>	
<b>Measures of association and agreement for describing land cover characterization classes .....</b>	<b>425</b>
<i>Eugene A. Fosnight and Gary W. Fowler</i>	
<b>Forest spatial surveys using the Rao-Hartley-Cochran sampling design .....</b>	<b>434</b>
<i>Jeffrey S. Pontius</i>	

(Continued)

<b>Increasing spatial precision and accuracy for monitoring peatlands in Switzerland by remote sensing techniques .....</b>	<b>441</b>
<i>Michael Köhl, Hans Jörg Schnellbacher, and Andreas Grünig</i>	
<b>Spatial and temporal models in contextual classification .....</b>	<b>451</b>
<i>Bo Ranneby</i>	
<b>Generalized linear mixed models for analyzing error in a satellite-based vegetation map of Utah .....</b>	<b>459</b>
<i>Gretchen G. Moisen, D. Richard Cutler, and Thomas C. Edwards, Jr.</i>	
<b>Statistical properties of measures of association and the kappa statistic for assessing the accuracy of remotely sensed data using double sampling .....</b>	<b>467</b>
<i>Mohammed A. Kalkhan, Robin M. Reich, and Raymond L. Czaplewski</i>	
<b>Comparing methods used to combine multi-level samples of compositional data ...</b>	<b>477</b>
<i>Murray Todd Williams</i>	
<b>Cost-effective, practical sampling strategies for accuracy assessment of large-area thematic maps .....</b>	<b>485</b>
<i>Stephen V. Stehman</i>	
<b>Spatial and probabilistic classification of forest structures using landsat TM data ...</b>	<b>493</b>
<i>Jeffrey L. Moffett and Julian Besag</i>	
<b>Total error estimation in a spatial database for GIS .....</b>	<b>501</b>
<i>Jose Alberto Quintanilha and Marcos Rodrigues</i>	
<b>Assessing the accuracy of a regional land cover classification .....</b>	<b>508</b>
<i>William Clerke, Raymond Czaplewski, Jeff Campbell, and Janet Fahringer</i>	
<b>Sampling satellite images for area estimates in a large region .....</b>	<b>509</b>
<i>F. J. Gallego</i>	
<b>Cooperative accuracy assessment strategies for sampling a natural landcover map of Arkansas .....</b>	<b>517</b>
<i>Robert S. Dzur, Michael E. Garner, Kimberly G. Smith, W. Fredrick Limp, Donald G. Catanzaro, and Richard L. Thompson</i>	
<b>Multidisciplinary</b>	
<b>Fuzzy measures in multi-criteria evaluation .....</b>	<b>527</b>
<i>J. Ronald Eastman and Hong Jiang</i>	
<b>On roles and goals for map accuracy assessment: a remote sensing perspective .....</b>	<b>535</b>
<i>Curtis E. Woodcock</i>	
<b>Integration of inventory and field data for automated fuzzy accuracy assessment of large scale remote-sensing derived vegetation maps in Region 5 national forests .....</b>	<b>541</b>
<i>Jeff A. Milliken and Curtis E. Woodcock</i>	
<b>Uncertainty of spatial metric relations in GIS .....</b>	<b>545</b>
<i>Xiaoyong Chen, Takeshi Doihara, and Mitsuru Nasu</i>	
<b>Plot collocation error: impacts on area estimation .....</b>	<b>553</b>
<i>Willem W. S. van Hees</i>	
<b>Sensitivity of landscape pattern metrics to misclassification and differences in land cover composition .....</b>	<b>560</b>
<i>James D. Wickham, Robert V. O'Neill, Kurt H. Riitters, Timothy G. Wade, and K. Bruce Jones</i>	
<b>Spatial (in)consistency of watershed delineations among agencies and scales in Pennsylvania .....</b>	<b>561</b>
<i>Wayne L. Myers, Barry M. Evans, and Michael C. Anderson</i>	

(Continued)

<b>Communicating the results of accuracy assessment: Metadata, digital libraries, and assessing fitness for use .....</b>	<b>568</b>
<i>Michael F. Goodchild</i>	
<b>Covariate-directed sampling for assessing species richness .....</b>	<b>569</b>
<i>G. P. Patil, Glen Johnson, and Matteo Grigoletto</i>	
<b>Data accuracy to data quality: using spatial statistics to predict the implications of spatial error in point data .....</b>	<b>577</b>
<i>Adam Lewis and Michael F. Hutchinson</i>	
<b>Moving into secondary map projections: an analysis of potential inconsistencies in spatial data .....</b>	<b>585</b>
<i>Mohamad N. Said and Peter F. Fisher</i>	
<b>Filling in missing forestry data: exploring autocorrelational techniques .....</b>	<b>593</b>
<i>Alissa N. Antle and Peter L. Marshall</i>	
<b>Optimizing sampling schemes for mapping and dredging polluted sediment layers .....</b>	<b>601</b>
<i>L. Hazelhoff and F. Hoefsloot</i>	
<b>Forest cover monitoring in India: the satellite experience .....</b>	<b>610</b>
<i>J. B. Lal</i>	
<b>Study about the use of SPOT data for estimation of pastoral areas in Morocco .....</b>	<b>617</b>
<i>Philippe Brion</i>	
<b>The influence of vegetation cover density and topographic parameters on the thermal emission of the beech forests of Simbruini Mountains (central Italy) .....</b>	<b>623</b>
<i>Carlo Ricotta, Giancarlo Avena, and Fernando Ferri</i>	
<b>Application of non-parametric kernel regression and nearest-neighbor regression for generalizing sample tree information .....</b>	<b>631</b>
<i>Annika Kangas and Kari T. Korhonen</i>	
<b>A method for assessing the prediction quality of mechanistic forest growth models .....</b>	<b>639</b>
<i>Biing T. Guan, George Gertner, and Pablo Parysow</i>	
<b>Statistical analysis of error propagation in national level carbon budgets .....</b>	<b>649</b>
<i>C. J. Cieszewski, D. P. Turner, and D. L. Phillips</i>	
<b>Monte Carlo simulations of nonlinear size-age relationships .....</b>	<b>659</b>
<i>Ronald E. McRoberts</i>	
<b>Cross-correlations among single tree growth models .....</b>	<b>667</b>
<i>Hubert Hasenauer, Robert A. Monserud, and Timothy G. Gregoire</i>	
<b>Genetics, geographics, and prairie dogs: error and accuracy in a validated spatially-explicit dispersal model .....</b>	<b>676</b>
<i>Gillian Bowser</i>	
<b>Effect of uncertainty in mapped biodiversity data on optimal conservation decisions .....</b>	<b>685</b>
<i>Michael J. Conroy and Jennifer E. Crocker</i>	
<b>Sensitivity analysis of species richness mapping to variations in forest-wildlife relationships .....</b>	<b>693</b>
<i>Tom Kohley</i>	
<b>Development of a survey sampling methodology for rare species .....</b>	<b>694</b>
<i>Molly Van Caster, David Bowden, and Jennifer Hoeting</i>	
<b>Methods to analyse the spatial structure of plant communities .....</b>	<b>695</b>
<i>Paul Braun, Heiko Baltzer, and Wolfgang Köhler</i>	
<b>Modeling population dynamics with cellular automata .....</b>	<b>703</b>
<i>Heiko Baltzer, Paul Braun, and Wolfgang Köhler</i>	

(Continued)

## Posters

<b>Inconsistency of line convolution as a quality factor in them aintenance of spatial databases .....</b>	<b>713</b>
<i>Michael J. C. Weir</i>	
<b>Lead-based paint survey data: spatial variation .....</b>	<b>714</b>
<i>Tze-San Lee</i>	
<b>Uncertainty of spatially-averaged rainfall estimates from rain gauges.....</b>	<b>715</b>
<i>Jeffrey R. McCollum and Witold F. Krajewski</i>	
<b>Defining urban settlements: certainties regarding uncertainties .....</b>	<b>716</b>
<i>Mahavir and Sahar Al-Amir</i>	
<b>Spatial dependence of tree biomass in an old-growth forest .....</b>	<b>717</b>
<i>Franco Biondi and Donald E. Myers</i>	
<b>Sensibility and uncertainty analyses of an expert system to determine stand treatments .....</b>	<b>718</b>
<i>Otto Eckmuellner and Martin Moser</i>	
<b>Stratified two-stage sampling (self-weighted) for assessment of village forest resources in Bangladesh.....</b>	<b>719</b>
<i>S. S. Islam</i>	
<b>Modelling animal movements within a GIS framework .....</b>	<b>720</b>
<i>J. Joachim, G. Janeau, and F. Spitz</i>	
<b>Evaluation of forest condition assessment data .....</b>	<b>721</b>
<i>F. Mitterböck and O. Eckmüllner</i>	
<b>Application of remote sensing technology in assessing village wood resources in Bangladesh .....</b>	<b>722</b>
<i>Jamil Ahmed Chowdhury</i>	
<b>Quality evaluation services on the Internet.....</b>	<b>723</b>
<i>Anders Östman</i>	
<b>Assessing attainment of standard by two stage sampling .....</b>	<b>724</b>
<i>Debashis Kushary</i>	
<b>Site positioning accuracies using GPS: Y-code receivers and real-time differential options .....</b>	<b>725</b>
<i>Karl E. Brown</i>	
<b>Statistical models of landscape pattern and the effects of coarse resolution of satellite imagery on estimation of area .....</b>	<b>726</b>
<i>Christine A. Hlavka</i>	
<b>Author index.....</b>	<b>727</b>

# Statistical Modeling of Environmental Data in Space and Time

Noel Cressie<sup>1</sup>

## EXTENDED ABSTRACT

As a concept, the environment is simply the surroundings of an organism or organisms. Space and time scales in environmental investigations can range from the very local to the very global. Some studies attempt to understand physical, chemical, and biological processes by performing controlled experiments in the laboratory. In this paper, we are concerned instead with studies made in the field. These are mostly observational in nature and, hence, even though a large amount of data may be collected and analyzed, one typically can only infer associations rather than causation.

Most environmental studies in the field involve variability over both space and time. The extension of traditional geostatistical methods, such as kriging, to the space-time domain is one possible approach to characterizing the variability of the processes (e.g., Bilonick, 1983; Eynon and Switzer, 1983; Stein, 1986; Le and Petkau, 1988; Sampson and Guttorp, 1992; Host et al., 1995). There are difficult modeling decisions to make in this approach, involving space, time, and space-time interaction components. The space-time variability is characterized by a variogram that often exhibits very different spatial behavior at different points in time and the class of variogram models that can be fit in this situation is very small indeed. In the atmospheric sciences, traditional methods for examining space-time processes have focused on Empirical Orthogonal Functions (EOF), Canonical Correlation Analysis (CCA), and Principal Oscillation Patterns (POP); see, for example, von Storch et al. (1995). Although these techniques are visually powerful, they were designed with summarization rather than prediction in mind.

Without the spatial component, there are a large class of time series that could be used to model the temporal component (e.g., autoregressive error processes). These are *dynamic* in the sense that they exploit the unidirectional flow of time. Without the temporal component, geostatistical methods could be used to model the spatial component (e.g., intrinsically stationary error processes). These are *descriptive* in the sense that although they model spatial correlation there is no causative interpretation associated with them. When both

---

<sup>1</sup> Noel Cressie is Professor of Statistics and Distinguished Professor in Liberal Arts and Sciences, Iowa State University, Ames, IA 50011.

temporal and spatial components are present, it seems sensible to use models that are a combination of both approaches, namely temporally dynamic *and* spatially descriptive. That is the new feature of our work and it allows a natural development of the space-time Kalman filter.

To give some definiteness to the problem, consider Markov temporal models with spatial colored noise:

$$S_t(\mathbf{s}) = \alpha_1 S_{t-1}(\mathbf{s}) + \alpha_2 S_{t-2}(\mathbf{s}) + \dots + \alpha_p S_{t-p}(\mathbf{s}) + \eta_t(\mathbf{s}); \mathbf{s} \in \mathbb{R}^2, \quad (1)$$

where  $S_t(\mathbf{s})$  is an (unobserved) value of the state process at location  $\mathbf{s}$  and time  $t$ . The observations are actually

$$Z_t(\mathbf{s}) = S_t(\mathbf{s}) + \epsilon_t(\mathbf{s}), \quad (2)$$

which expresses the data as a noisy version of the state process. The goal is to predict  $\hat{S}_{t_0}(\mathbf{s}_0)$ , where both  $t_0$  and  $\mathbf{s}_0$  may or may not represent space-time coordinates at which data are available. In (1),  $(\alpha_1, \dots, \alpha_p)$  are autoregressive parameters (i.e., parameters of the "temporally dynamic" component) and, most importantly,  $\eta_t(\cdot)$  is a spatially-colored noise process (i.e., the "spatially descriptive" component). In (2),  $\epsilon_t(\cdot)$  is a white-noise process representing measurement error.

Data come in the form,

$$\mathbf{Z}_j \equiv (Z_j(\mathbf{s}_1), \dots, Z_j(\mathbf{s}_n))'; j = 1, \dots, t,$$

where it is not essential that all observations are available at each time point and at each spatial location. The optimal predictor of  $S_{t_0}(\mathbf{s}_0)$  is:

$$\hat{S}_{t_0|t}(\mathbf{s}_0) \equiv E[S_{t_0}(\mathbf{s}) | \mathbf{Z}_1, \dots, \mathbf{Z}_t],$$

with mean-squared prediction error,

$$E(\hat{S}_{t_0|t}(\mathbf{s}_0) - S_{t_0}(\mathbf{s}_0))^2.$$

Both these quantities can be calculated recursively, using what could be called a *space-time Kalman filter* (Huang and Cressie, 1996).

The state-space model in (1) is very attractive because it features the dynamic aspect through an autoregressive structure but builds in space-time interaction through the error process  $\eta_t$ , which is, at any point in time, a spatially correlated (e.g., intrinsically stationary) process. Notice that  $S_t(\mathbf{s})$  is influenced directly by past values only at location  $\mathbf{s}$ . In reality, spatio-temporal processes are likely to be more complicated, to have dependence also on past values at locations  $\mathbf{u}$  *near*  $\mathbf{s}$ . Thus, I shall investigate the spatio-temporal climate model,

$$S_t(\mathbf{s}) = \alpha \int \omega_s(\mathbf{u})S_{t-1}(\mathbf{u})d\mathbf{u} + \eta_t(\mathbf{s}), \quad (3)$$

where, for identifiability, the coefficients  $\omega_s(\mathbf{u})$  satisfy  $\int \omega_s(\mathbf{u})d\mathbf{u} = 1$ . Development of the spatio-temporal Kalman Filter allows prediction of  $S_t(\mathbf{s}_0)$  based on data  $\mathbf{Z}_1, \dots, \mathbf{Z}_t$ .

The two challenging aspects of this research are to derive the expressions for the optimal predictor and the optimal mean-squared prediction error, and to obtain efficient estimators of the parameters  $\alpha$ ,  $\text{var}(\epsilon_t(\mathbf{s}))$ , and  $\text{cov}(\eta_t(\mathbf{s}), \eta_t(\mathbf{u}))$ . These are then substituted into the optimal prediction equations. (This research is in progress with Ph.D. student Christopher K. Wikle.) An alternative to the estimation of parameters is to put (prior) distributions on them. In this case, there is a good physical reason to do this. The parameters almost certainly vary from year to year and from region to region; this extra variation can quite simply be handled by replacing  $\alpha$  in (3) with  $\{\alpha_0, \dots, \alpha_{t-1}, \alpha_t, \dots\}$  and assuming them to be distributed according to some prior distribution. The goal is still to obtain  $E(S_t(\mathbf{s}_0) | \mathbf{Z}_1, \dots, \mathbf{Z}_t)$  and its mean-squared prediction error or, more generally, the posterior distribution of  $S_t(\mathbf{s}_0)$  given  $\mathbf{Z}_1, \dots, \mathbf{Z}_t$ . While such calculations were daunting five years ago, Markov chain Monte Carlo (e.g., Bernardo and Smith, 1994, Section 5.5.5) can be invoked to handle the problem.

#### REFERENCES

- Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory*. Wiley, Chichester.
- Bilonick, R.A. (1983). Risk qualified maps of hydrogen ion concentration for the New York state area for 1966-1978. *Atmospheric Environment*, **17**, 2513-2524.
- Eynon, B.P. and Switzer, P. (1983). The variability of rainfall acidity. *Canadian Journal of Statistics*, **11**, 11-24.
- Host, G., Omre, H., and Switzer, P. (1995). Spatial interpolation errors for monitoring data. *Journal of the American Statistical Association*, **90**, 853-861.
- Huang, H.C. and Cressie, N. (1996). Spatio-temporal prediction of snow water equivalent using the Kalman filter. *Computational Statistics and Data Analysis*, in press.
- Le, D.N. and Petkau, A.J. (1988). The variability of rainfall acidity revisited. *Canadian Journal of Statistics*, **16**, 15-38.
- Sampson, P.D. and Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, **87**, 108-119.
- Stein, M.L. (1986). A simple model for spatial temporal processes. *Water Resources Research*, **22**, 2107-2110.
- von Storch, H., Burger, G., Schnur, R., and von Storch, J.S. (1995). Principal oscillation patterns: A review. *Journal of Climate*, **8**, 377-400.

# A Case Study in Geostatistical Modeling for Petroleum Reservoir Description

Jeffrey M. Yarus<sup>1</sup>, Jeffrey A. May<sup>2</sup>, and Timothy C. Coburn<sup>3</sup>

---

**Abstract.**--Geostatistical techniques can be used to effectively characterize the subsurface properties of petroleum reservoirs, even in the face of limited amounts of control data. The applications described here pertain to a newly-discovered offshore oil field in southeast Asia, the extent of which covers approximately 35 square kilometers. The study involves the development of a three-dimensional grid of estimated rock properties (e.g., porosity and lithology), consisting of approximately 3.1 million nodes, to be used in computer simulations of field production performance. A project of this scope requires a novel combination of modeling procedures and spatial analysis tools, including embedded Markov chain analysis, sequential Gaussian simulation, and probability field simulation. A description of the overall approach is presented, along with an example of the modeling results.

---

## INTRODUCTION

Petroleum engineers use computer simulations to establish the likely production profiles of oil and gas fields and reservoirs. Such simulations are based on a number of input parameters which are estimated through empirical observation, direct and indirect measurement in the subsurface, or by analogy to similar hydrocarbon bearing provinces. The resulting profiles are used to guide further development of fields through drilling of additional wells or by the adoption of enhanced extraction or recovery methods.

The data needed to estimate the input parameters for the computer simulations includes both qualitative and quantitative descriptions of the geological regime, downhole measurements (acoustic, electrical, etc.) of the subsurface media, laboratory measurements on soil, rock, and fluid samples, seismic reflections, drilling records, and subjective geological experience and intuition. All too often,

---

<sup>1</sup> *Consulting Mathematical Geologist, Yarus & Associates, Denver, CO.*

<sup>2</sup> *Geological and Geophysical Consultant, Littleton, CO.*

<sup>3</sup> *Senior Statistician, National Renewable Energy Laboratory, Golden, CO.*

however, many of the requisite data are missing or simply unavailable, a situation which is particularly true in the case of newly-discovered fields, or in foreign and/or offshore settings, where data collection can be a comparatively expensive and logistically difficult proposition.

Historically, the two most common ways to remedy missing data have been to impute it from knowledge of other producing fields in geologically similar environments (best case) or to simply “make it up” (worst case). As an alternative, the discipline and tools of geostatistics have recently been shown to provide a more rigorous and statistically-reliable means of estimating the simulation parameters by taking advantage of the notion of spatial correlation that is so central to the investigation of all earth processes and phenomena.

This paper discusses the effectiveness of the geostatistical approach in a laterally extensive geographic setting, and with the additional limitation of minimal control points and other tangible data. A number of spatial analysis tools, including embedded Markov chain analysis, conditional simulation, and probability field simulation are combined in novel ways to produce a three-dimensional grid of parameter estimates encompassing approximately 3.1 million nodes. A description of the overall modeling approach is presented, along with an example of the modeling results.

## **STUDY SETTING**

The present study involves a newly-discovered offshore oil field in southeast Asia for which a plan of development is necessary to satisfy the contractual requirements of the business concession. Three successful wells (A, B, and C) have been drilled. Drilling records, along with preliminary seismic data and initial production tests on the wells, confirm the presence of hydrocarbons in an area covering approximately 35 square kilometers. Because of the geographical extent of the field, and the high cost of drilling additional wells, it is desirable to develop the most precise model of the reservoir possible, prior to selecting additional drilling sites or production platform locations. Such a model must reliably characterize the degree and lateral extent of all reservoir properties which affect fluid flow and hydrocarbon production.

### **Conceptual Geological Framework**

Preliminary regional studies of the field area identified the presence of an ancient rift system filled with clastic sediment, sourcing from the northeast. The

hydrocarbon-bearing interval (the reservoir) of the field is located in one particular subsurface horizon, which was apparently deposited early in the rift fill sequence. Using this information, a conceptual model was developed which establishes the depositional environment as being a fluvial dominated, fan delta system.

### Available Data

The data set consists of ten seismic lines from an early 2-D survey over the region, and information from three wells drilled along a general northwest-to-southeast transect in the central southeast quadrant of the study area, the wells being spaced at intervals of approximately 2000 meters. During the drilling operation, cores were cut from two of the wells and sent to the laboratory to support rock studies. Conventional suites of wireline logs (including sonic, density, resistivity, neutron, and gamma ray) were run in all three wells.

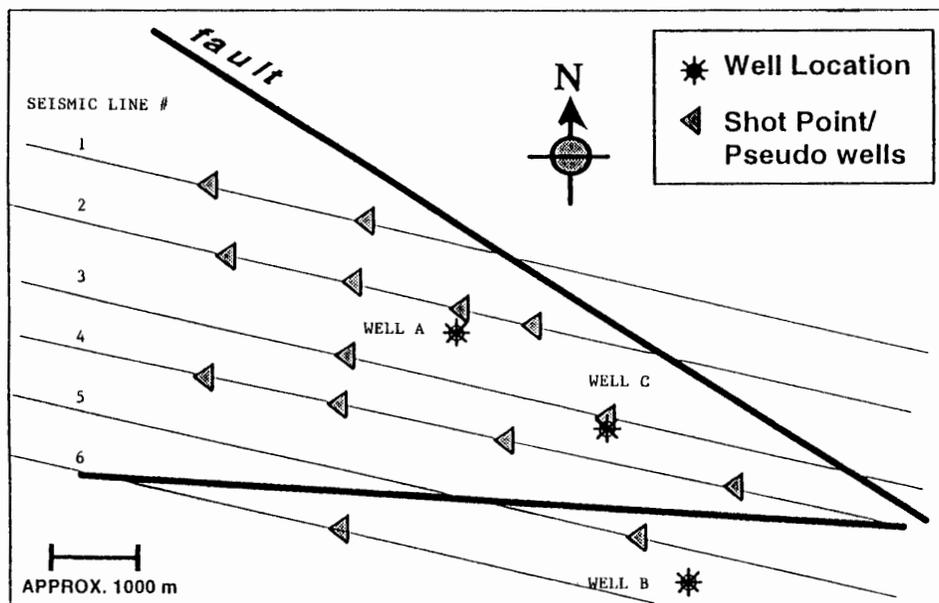


Figure 1.--Relative locations of wells, pseudo wells, seismic lines, and faults

Values of reservoir parameters at the well locations were determined from laboratory measurements and examination of the core material, as well as from numerical analysis of the well logs. In particular, individual values of porosity were obtained for each half-foot segment of the reservoir. Similarly, each half-foot segment was classified into one of four primary lithologies, or rock types--channel sands, delta front sands, overbank sands and muds, and lacustrine muds--identified by the project geologists. Of these four lithologies, the channel sands are presumed to have the greatest potential for production. The resulting collection of

porosity measurements and lithology classifications comprises approximately 2,500 data pairs per well.

Based on further studies of the cores, analysis of the wireline logging runs, and interpretation of the seismic lines, the following additional conclusions were formulated: (1) the field is bounded by at least two major faults, with Well B being situated outside these boundaries (Figure 1); (2) drainage is controlled, in part, by east-west trending faults; (3) four major depositional cycles (Intervals 1 to 4), separated by non-permeable shale barriers, can be identified which effectively subdivide the total reservoir into separate producing units of varying thickness; (4) the principle orientation of the channel sands, the dominant reservoir lithology, is also east to west; (5) the geometry of the channel sands is represented by an average length-to-width ratio of approximately 10:1 (taken in part from results reported by Fielding and Crane (1987) and Knutson and Boardman (1978)); and (6) the median estimated thickness of the channel sands is 6 ft.

#### **EXPANDING AERIAL DATA COVERAGE THROUGH PSEUDO WELL DEVELOPMENT**

By themselves, the three wells do not constitute an adequate geographic sampling of the reservoir from which to develop a three-dimensional model. Therefore, in order to increase the aerial distribution of "sampling" locations and to facilitate geostatistical interpolation, fourteen pseudo wells were introduced. Using the process described below, these pseudo wells were created in such a way as to reflect the nature of the reservoir observed at the actual drillsite locations, adjusted for the features of the conceptual geological framework, relationships extracted from geological outcrop studies, and trends observed in the seismic data.

Pseudo well sites were determined to coincide with several of the shot point locations identified along the ten seismic lines, resulting in the "sampling" coverage illustrated in Figure 1. The shot point locations provided a convenient additional data source in that the seismic velocity recorded at these sites could be used to calculate acoustic impedance (AI), a continuous reservoir property from which the boundaries of the four major depositional cycles could be estimated. To do so, log suites and synthetic seismograms from the three wells were manually depth-shifted to obtain an optimum visual alignment of all petrophysical markers. This identification of the major cycle boundaries was corroborated through an automated "correlation" approach (Olea, 1988). The depths associated with these boundaries were then projected to the acoustic impedance trace at each pseudo well location.

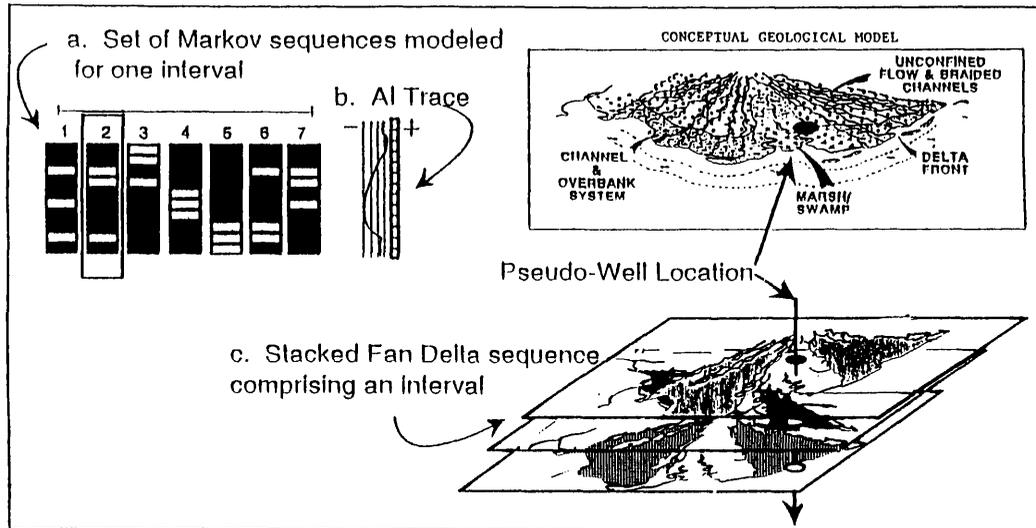


Figure 2.--The process for determining lithology in pseudo wells

### Markov Modeling of Lithology

The second stage of pseudo well construction involved overprinting a vertical sequence of lithologies at each location. Embedded Markov chain analysis (Davis, 1986) was used to develop these sequences based on the transition probabilities (bottom to top) observed in the actual wells. For illustration purposes, Table 1 shows the transition probabilities for Interval 1 computed over the three wells.

Table 1.--Interval 1 transition probabilities (bottom to top) for the four lithologies

		To			
		Channel Sands	Delta Front Sands	Overbank Sands & Muds	Lacustrine Muds
From	Channel Sands	0	0	.78	.22
	Delta Front Sands	.28	0	0	.72
	Overbank Sands & Muds	.67	0	0	.33
	Lacustrine Muds	.19	.82	0	0

Applying the marginal vectors, seven different lithologic sequences were generated

for each of the four intervals at each pseudo well location. Stacking the interval sequences together resulted in 28 possibilities per pseudo well, from which the most likely case was selected by subjective comparison to the respective acoustic impedance depth traces, giving due consideration to the conceptual geological model. The result of this effort was the assignment of a lithology classification to each half-foot depth interval in each of the pseudo wells (Figure 2).

### **Determination of Porosity Via Sequential Gaussian Simulation**

The third stage of pseudo well construction involved determination of a value of porosity for each half-foot depth interval. This task required two critical inputs: (1) the cumulative distribution (CDF) of porosity for each lithology within each of the four major intervals, combining information across wells; and (2) a vertical correlogram of porosity for each of the four intervals (combining information across wells and ignoring lithology). The CDFs were used to guarantee that the statistical distribution of porosity values in the pseudo wells matched that of porosity values in the original wells; and the correlogram, which measures the degree to which porosity changes from one depth to another, was used to insure that the vertical spatial distributions of porosity in the pseudo wells and the original wells were aligned.

Given the CDFs and correlograms described above, porosity values for each individual depth in each pseudo well were determined using sequential Gaussian simulation (Deutsch and Journel, 1992), a form of conditional simulation, in the following way. For each pseudo well, a vertical depth was selected at random from its vertical section. The interval and lithology associated with that depth were identified, and a porosity value was then randomly selected employing the appropriate CDF. This Monte-Carlo-like approach was successively repeated for every depth in the selected well. When a selected depth was located near one that had been previously simulated (i.e., within the range of correlation defined by the appropriate correlogram), and the two depths were associated with the same lithology, the value of porosity at the first depth influenced determination of the value at the second. The degree of influence exerted was dependent on the correlation coefficient defined by the correlogram for the distance (in depth) separating the two.

### **3-D PROPAGATION OF LITHOLOGY AND POROSITY USING PROBABILITY FIELD SIMULATION**

Using information from the three wells and the fourteen, fully-populated pseudo wells, a full three-dimensional model of the reservoir was developed. The initial

step was to construct a rectangular three-dimensional network grid consisting of approximately 3.1 million nodes. Grid dimensions were set at  $180_x \times 87_y \times 200_z$  cells, with individual cell dimensions being  $x = 150$  ft,  $y = 150$  ft, and  $z = 5$  ft (the  $z$ -dimension represents the thickness of one reservoir layer). A grid of this density was deemed necessary to depict the detailed lithologic changes characteristic of this reservoir.

Probability field simulation (Srivastava, 1992) was used to propagate porosity values and lithology classifications from the wells and pseudo wells to the individual grid node locations. Separate simulations were conducted for the two quantities, the results of which were subsequently combined for purposes of describing the reservoir in a multivariate sense.

Like sequential Gaussian simulation, the use of probability field simulation (another form of the more general technique of conditional simulation) required information about the statistical distribution of the reservoir properties, as characterized by their CDFs, as well as spatial variation in both the vertical and lateral directions, as characterized by the appropriate correlograms. Because lithology is not continuous, CDFs and directional correlograms for this quantity were determined using an indicator variable (having values 1 or 0) which designates the presence or absence of the specific lithology in question.

In the case of porosity, spatial variation was represented in the vertical sense using the same correlogram (described above) used to establish porosity values in the pseudo wells. To characterize the spatial variation in the lateral sense, additional directional correlograms for each major interval were constructed using the porosity values available in all three wells and fourteen pseudo wells. Similarly, both vertical and lateral correlograms were developed for each lithology within each major interval (combining information across all wells and pseudo wells) based on the assigned indicator values.

The application of probability field simulation, the algorithm for which is described more fully by Srivastava (1992), required the development of a three-dimensional grid of randomly-assigned probabilities (a probability template) based on both the vertical and lateral correlograms of the quantity of interest. This grid had the same physical dimensions as the simulation grid, with each node of the probability template corresponding to a node on that grid. The probability values assigned to the nodes were then used to select values of porosity or lithology from the appropriate CDFs, resulting in a fully-populated simulation grid.

## EXAMPLE OF THE SIMULATION RESULTS

Based on the dimensions of the simulation grid, the reservoir was arbitrarily subdivided into 200 five-foot layers to optimize precision and to better take advantage of the available computing resources. All results were generated on Unix-based workstations, using a combination of proprietary and commercially-available code.

Individual simulations of both porosity and lithology were developed for each of the 200 layers, with several realizations of each layer being produced. Figure 3 graphically illustrates the results for one layer of porosity, depicting a lateral gradational character across the reservoir which closely mimics input from the conceptual geological model, along with the well data and seismic interpretations.

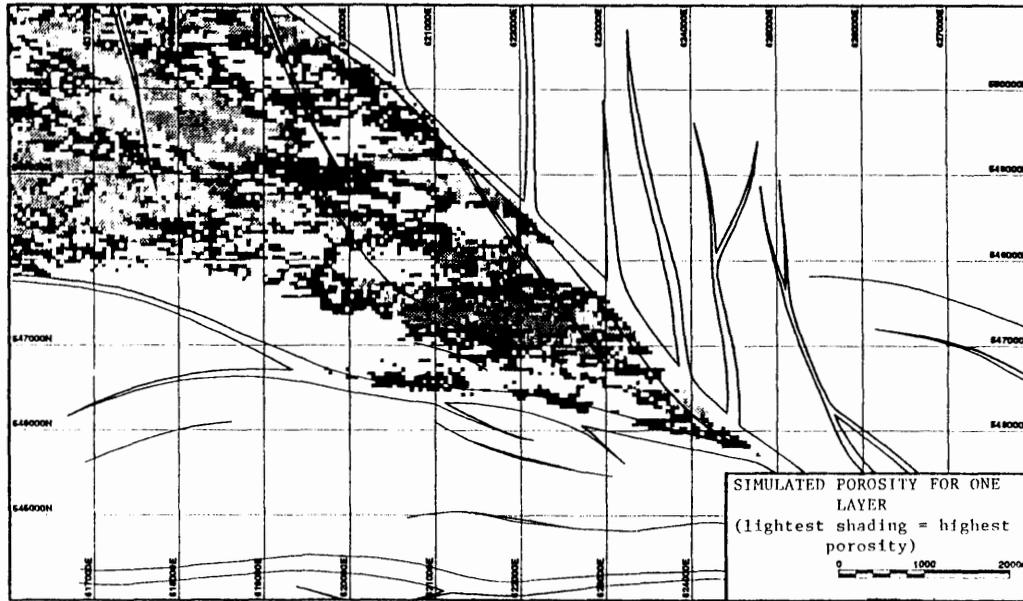


Figure 3.—Simulation results for porosity in one reservoir layer

## SUMMARY

A number of geostatistical techniques and other quantitative methods, including conditional simulation, probability field simulation, and Markov analysis, were combined to fully characterize and model a large-scale oil reservoir. Drawing on concomitant information from a number of sources, the unique process that was employed allowed the complicating situation of a limited number of control points (wells) to be overcome. The study resulted in reservoir models which exactly honored the control points, and which corroborated the field production profiles of other independent engineering tests. The geostatistical approach, however,

yielded results having much more lateral and vertical detail which could be used to more effectively guide the placement of future additional wells and/or drilling platforms.

## REFERENCES

- Davis, John C. 1986. *Statistics and data analysis in geology*, 2nd ed. New York: John Wiley & Sons. 646 p.
- Deutsch, C. V. and A. G. Journel. 1992. *GSLIB--geostatistical software library and user's guide*. New York: Oxford University Press. 2 diskettes and 340 p.
- Fielding, C. R. and R. C. Crane. 1987. An application of statistical modeling to the prediction of hydrocarbon recovery factors in fluvial reservoir sequences, in *Recent Development in Fluvial Sedimentology* (F. G. Ethridge, R. M. Flores, and M. D. Harvey, eds.). Society of Economic Paleontology and Mineralogy, Special Publication 39, p. 321-327.
- Knutson, C. F. and C. R. Boardman. 1978. Continuity and permeability development in the tight gas sands of the eastern Uinta Basin, Utah. *Proceedings of the Fourth DOE Symposium on Enhanced Oil & Gas Recovery and Improved Drilling Methods*, v. 2, p. E-2/1-24.
- Olea, Ricardo A. 1988. *CORRELATOR--An interactive computer system for lithostratigraphic correlation of wireline logs*. Lawrence, KS: Kansas Geological Survey, Petrophysical Series 4. 84 p.
- Srivastava, R. M. 1992. Reservoir characterization with probability field simulation. *Proceedings of the 67th Annual Technical Conference and Exhibition of Society of Petroleum Engineers*, SPE Publication 24753, p. 927-938.

## BIOGRAPHICAL SKETCH

Jeffrey M. Yarus is a consulting mathematical geologist and principal in the firm Yarus & Associates in Denver, Colorado. He holds a Ph.D. in geology from the University of South Carolina, and specializes in applications of geostatistics in petroleum reservoir characterization, geoscientific computer mapping, and geostatistical software development.

Jeffrey A. May is an independent consulting geologist based in Littleton, Colorado. He holds a Ph.D. in geology from Rice University, and specializes in sedimentology, seismic interpretation, and sequence stratigraphy.

Timothy C. Coburn is senior statistician at the National Renewable Energy Laboratory in Golden, Colorado. He holds a Ph.D. in statistics from Oklahoma State University, and specializes in the analysis and modeling of energy and environmental data, with particular emphasis on geoscience investigations.

# An Overview of Stochastic Spatial Simulation

R. Mohan Srivastava<sup>1</sup>

---

**Abstract.** — Though tools for stochastic spatial simulation are becoming increasingly available and accessible, the variety of different algorithms also forces users to make conscious choices about the methodologies they select. This overview paper presents a brief summary of two of the most successful and popular approaches, sequential simulation and simulated annealing, and uses these as the basis for discussing some of the practical and theoretical issues surrounding the use of stochastic spatial simulation.

---

## INTRODUCTION

In a broad range of earth science applications, stochastic spatial simulation is rapidly becoming a common tool for engineering and decision-making in the face of uncertainty. In environmental applications, it is being used as the basis for remediation planning and confirmation sampling, not only at large Superfund sites but also for small backyard contamination problems. In fisheries studies, stochastic spatial simulation is being used by regulatory agencies to improve their ability to manage fish stocks and to guarantee the stability of the fisheries industry. In agricultural studies, the same tools are being used to determine whether or not policies designed to improve soil properties are having their intended effect. As the interest in stochastic spatial simulation grows, there is an explosion in the variety of simulation algorithms. Even a comprehensive toolkit like the public domain GSLIB software (Deutsch and Journel, 1992), with its 11 routines for simulation, does not cover the breadth of what is currently available. With simulation research thriving at several academic institutions, broad choice is starting to become bewildering confusion.

## WHAT DO WE WANT FROM A STOCHASTIC METHOD?

### Stochastic models as art

Stochastic models often are not used quantitatively and serve only as colorful wallpaper for the halls of research centers. Though such artwork is often dismissed as lacking in substance, it can play a useful role in catalyzing better technical work. In industries where there is an endemic belief that all spatial models should look like contour maps, it is difficult to get people to appreciate that there is much more complexity between the available sample data than

---

<sup>1</sup>Geostatistician, FSS International, Vancouver, B.C., Canada

traditional contouring techniques can portray. Stochastic models often have their greatest influence when they are used to challenge a complacent belief in the simplicity of spatial processes. When engineers and scientists who are familiar with traditional contouring techniques first see the results of a stochastic approach, their initial reaction is often one of surprise and skepticism. Their surprise is due to the fact that stochastic models are much more visually complex than the models they are familiar with; their skepticism is due to the fact that there are many different renditions, rather than a unique “best” model. This initial surprise and skepticism often gives way to curiosity, however, as they realize that these alternate renditions all share a similar visual character and all somehow manage to honor the available data.

Once a stochastic model is rendered on maps and cross-sections and hung on the walls of a meeting room, it often becomes a lightning rod for criticism as people identify specific features that cannot possibly exist. Their reasons vary from geologic arguments based on depositional environment to geophysical arguments based on surface seismic studies to engineering arguments based on pressure tests to biologic arguments based simple principles of food supply. In all of these situations, the benefit of the stochastic model may simply be to focus the views of a wide variety of experts and, in so doing, to point the direction to an improved model. Whether this new and clearer understanding is used to improve the stochastic model or is used to make intelligent adaptations to more traditional models, the initial stochastic model has still played an important role in the evolution of an appropriate spatial model.

### **Stochastic models for assessing the impact of uncertainty**

Many of the engineers and scientists who are responsible for forecasting the future behavior of a spatial process, such as fluid flow in a reservoir or contaminant transport in an aquifer, have a solid understanding of the fact that there is always uncertainty in any spatial model. They know that their forecasts are based on a specific model that is designated the “best” model, but they also want a couple of other models, a “pessimistic” one and an “optimistic” one, so that they can assess whether the decisions and policies that have been developed on the “best” model are flexible enough to handle the uncertainty.

When used for this kind of study, a stochastic approach offers a variety of spatial models, each one of which is consistent with the available data. We can sift through these many possible renditions and select one that looks pessimistic and another that looks optimistic.

### **Stochastic models for Monte Carlo risk analysis**

Stochastic modelling offers the hope of supporting a Monte Carlo type of risk analysis since the various renditions it produces are not only plausible, in the

sense that they honor all the available data, but they are also intended to be equally probable in the sense that any one of them is as likely a representation of the true underlying spatial phenomenon as any other one. In such studies, hundreds of alternate models, if not thousands, are generated and processed to produce a distribution of possible values for some critical engineering parameter — breakthrough time of a contaminant, for example, or an economically optimal ore/waste classification; these distributions are then used to optimize some decision by minimizing an objective function.

A critical aspect of this use of stochastic modelling is the belief in a “space of uncertainty” and that the stochastic modelling technique can produce outcomes that sample this space fairly. We know that we cannot possibly look at *all* possible outcomes, but we believe that we can get a fair representation of the whole spectrum of possibilities. When used for this purpose, we hope that our stochastic modelling technique does not have any systematic tendency to show us optimistic or pessimistic scenarios since we are going to use all of the outcomes as equally probable representations of reality.

While these types of studies overlap somewhat with those described in the previous section, the important difference between the two is that Monte Carlo risk analysis involves the notion of a probability distribution while the type of study described in the previous section does not. The previous section referred to studies where models are chosen by sifting through a large set and extracting two that seem to be plausible, but extreme, scenarios. No-one particularly cares whether the “optimistic” case is the 95th percentile or the 99th percentile; it’s just an example of a case that produces a very optimistic result while still honoring the same basic data. In Monte Carlo risk analysis, however, we depend on the notion of a complete probability distribution of possible outcomes and on the hope that our stochastic approach is producing outcomes that fairly represent the entire space.

### **Stochastic models for honoring spatial variation**

Though stochastic techniques can produce many outcomes, stochastic modelling studies often use a single outcome as the basis for forecasting and planning. When used in this way, spatial simulation techniques are attractive not so much for their ability to generate many outcomes but rather for their ability to produce outcomes that have realistic spatial variation.

Over the past decade it has become increasingly clear that forecasts of behavior of spatial processes are more accurate when based on models that reflect the actual spatial heterogeneity of the phenomenon under study. In the petroleum industry, there are countless examples of reservoir performance predictions whose failure is due to the use of overly simplistic models; similar disasters can be found in environmental remediation plans that were based on smooth

contour maps that could not possibly be used to assess the effect of unexpected “hot spots”. Most traditional methods for modelling spatial processes end up with a model that is too smooth and continuous, gently undulating from one sample location to the next rather than showing the spatial variation that is known to exist between samples. Such smoothness often leads to biased predictions and poor decisions — contaminants escape from their repositories much quicker than expected, for example, or the actual volumes of remediated soil end up being much larger than originally forecast.

Though using a stochastic method to produce a single outcome is often viewed with disdain by those who prefer to pump hundreds of possible outcomes through Monte Carlo risk analysis, it can be argued that even a single outcome from a stochastic approach is a better basis for prediction and decision-making than a single outcome from a technique that does not honor spatial variation.

## TWO COMMON STOCHASTIC METHODS

### Sequential simulation

The family of “sequential” procedures all make use of the same basic algorithm:

1. Choose at random a node where no simulated value exists.
2. Estimate the local conditional probability distribution (lcpd).
3. Draw at random a single value from the lcpd.
4. Include the newly simulated value in the set of conditioning data.
5. Repeat steps 1 through 4 until all grid nodes have a simulated value.

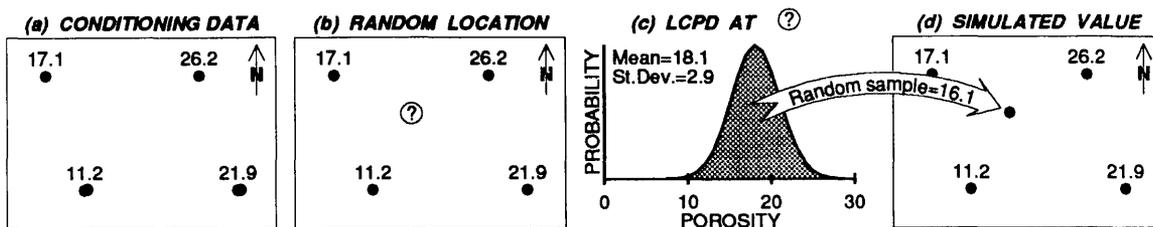
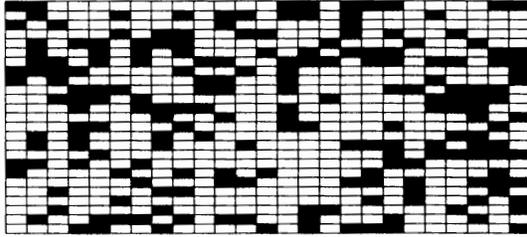


Figure 1: Sequential simulation of porosities in an aquifer.

The only significant difference between the various sequential procedures is estimation of the lcpd; any technique that estimates the lcpd can be used to drive sequential simulation. Multigaussian kriging, for example, produces an estimate of the lcpd by assuming that it follows the normal distribution and estimating its mean and standard deviation; this was the approach used in the diagram in Figure 1. Indicator kriging is an example of another technique that could be used to estimate the lcpd. With this procedure, no assumption is made about the shape of the distribution and the lcpd is constructed by directly estimating the probability of being below a series of thresholds or the probability of being within a set of discrete classes.

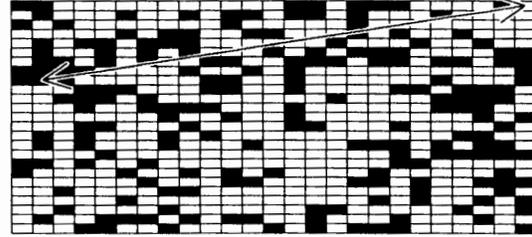
**(a) INITIAL GRID**  
 Note: Each pixel is 20m horizontally by 3m vertically  
 Net-to-gross: 69.7%  
 Ave. length: 29.78m  
 Ave. thickness: 4.53m

Actual Target  
 69.7% 70%  
 29.78m 60m  
 4.53m 10m



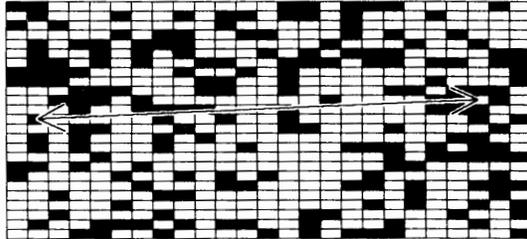
**(b) ITERATION NO. 1**  
 (2,17) ↔ (25,25)  
 Statistics improve,  
 keep this grid  
 Net-to-gross: 69.7%  
 Ave. length: 30.00m  
 Ave. thickness: 4.57m

Actual Target  
 69.7% 70%  
 30.00m 60m  
 4.57m 10m



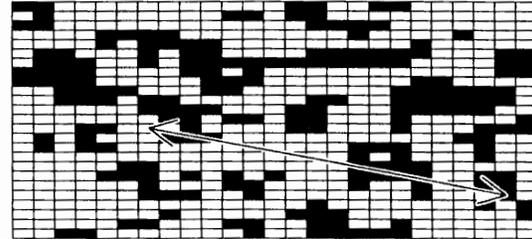
**(c) ITERATION NO. 2**  
 (2,13) ↔ (23,15)  
 Statistics deteriorate,  
 go back to last grid  
 Net-to-gross: 69.7%  
 Ave. length: 29.34m  
 Ave. thickness: 4.45m

Actual Target  
 69.7% 70%  
 29.34m 60m  
 4.45m 10m



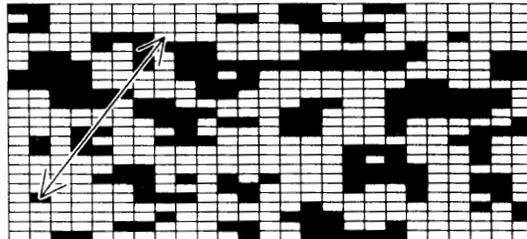
**(d) ITERATION NO. 463**  
 (24,5) ↔ (7,12)  
 Statistics improve,  
 keep this grid  
 Net-to-gross: 69.7%  
 Ave. length: 45.68m  
 Ave. thickness: 6.55m

Actual Target  
 69.7% 70%  
 45.68m 60m  
 6.55m 10m



**(e) ITERATION NO. 464**  
 (2,5) ↔ (8,22)  
 Statistics deteriorate,  
 go back to last grid  
 Net-to-gross: 69.7%  
 Ave. length: 45.46m  
 Ave. thickness: 6.51m

Actual Target  
 69.7% 70%  
 45.46m 60m  
 6.51m 10m



**(f) ITERATION NO. 1987**  
 (25,14) ↔ (14,12)  
 Statistics improve,  
 keep this grid  
 Net-to-gross: 69.7%  
 Ave. length: 60.00m  
 Ave. thickness: 9.88m

Actual Target  
 69.7% 70%  
 60.00m 60m  
 9.88m 10m

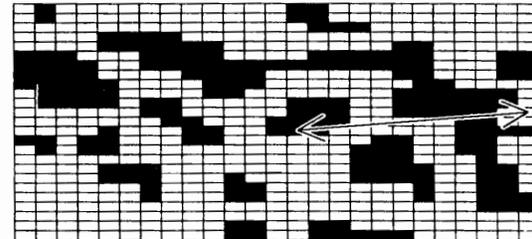


Figure 2: Simulated annealing for a simple sand/shale simulation.

### Simulated annealing

In simulated annealing the spatial model is constructed by iterative trial-and-error. Figure 2 shows a simplified annealing procedure for the problem of producing a sand/shale model that has a 70% sand, an average shale length of 60 meters and an average thickness of 10 meters. In Figure 2a the pixels in the model are randomly initialized with sands and shales in the correct global proportion. The random assignment of the sands and shales causes the average shale length and thickness to be too short. Figure 2b shows what happens after we perturb the initial model by swapping a sand pixel with a shale pixel. Such a swap will not affect the global proportions of sand and shale; it does, however, make a tiny improvement in both the average shale length and thickness. Figure 2c shows the grid after a second swap. We're not

quite as lucky with this second swap since it makes the average shale length and thickness a little worse, so we reject this second swap, go back to the previous grid shown in Figure 2b and try again.

This swapping is repeated many more times and, after each swap, the average shale length and thickness are checked to see if we are any closer to the target values. If a particular swap does happen to get us closer to the length and thickness statistics we want then we keep the swap; on the other hand, if the swap causes the image to deteriorate (in the sense that the length and thickness statistics are further from our target) then we undo the swap, go back to the previous grid and try again.

Figure 2d shows the grid after the 463rd swap, which happens to be a good one. At this point, 66 of the attempted 463 swaps were good and were kept, while the remaining 397 were bad and were not kept. Figure 2e shows the effect of the 464th swap, a bad one that is not kept because it causes the statistics to deteriorate. Figure 2f shows what the grid looks like after the 1987th swap. At this point there were 108 good swaps and 1879 bad ones, and the shale length and thickness statistics are very close to our target values.

The key to simulated annealing is minimizing the deviation between the grid statistics and target values; the function that describes this deviation is usually called the “energy” or “objective” function. In the example in Figure 2, where we were trying to match the average length and thickness of the shales, the energy function was simply the sum of the absolute differences between the actual average dimensions and our target values:  $E = |\text{Actual average length} - 60| + |\text{Actual average thickness} - 10|$ . The annealing procedure tries to drive the value of  $E$  to zero; decreases in energy are good, increases are bad.

Though Figure 2 displays some of the key concepts in simulated annealing, it is a very simplified version and does not show some of the important aspects of practical implementations of simulated annealing. Most simulated annealing procedures do not recalculate the energy after each swap. What they do instead is perform several swaps before recalculating the energy. Initially, when the energy is high (the statistics of the grid are far away from the target values) many swaps are performed before the energy is recalculated; as the energy decreases and the grid statistics converge to their desired target values, the number of swaps between recalculations is decreased.

The second important difference between most practical implementations of annealing and the simplified example in Figure 2 is tolerance of lack of progress. In the example shown in Figure 2, any swap that increased the energy (moved us further away from our target statistics) was immediately undone. In the common practical implementations of simulated annealing, some increases in energy are tolerated. The chance of allowing the energy to increase is related to two factors: the magnitude of the increase and the number of iterations. In-

creases in energy are more tolerated early on in the procedure; small increases are always more tolerated than large ones.

A third difference between the example given here and most practical implementations is that the example in Figure 2 does not have any conditioning information to honor. Conditioning information is quite easily honored in practice by setting the appropriate pixels to the values observed and then ensuring that these conditioning pixels are never swapped later in the procedure.

The other important difference between most practical implementations of simulated annealing and the simplified example in Figure 2 is that the energy function is usually much more complex than the simple one used here. Simulated annealing is a very flexible approach that can produce stochastic models that honor many different kinds of information; the paper by Kelkar and Shibli in this volume provides an example of the use of annealing for honoring the fractal dimension. For any desired features or properties that we can summarize numerically, we can use simulated annealing to honor these features by including appropriate terms in the energy function. In Figure 2, for example, we wanted to honor both the average shale thickness and the average shale length so our energy function had two components. In more interesting practical examples, the energy function might have several more components. For example, we could have one component that describes how close the grid's variogram comes to some target variogram, a second component that describes how similar the synthetic seismic response of the grid comes to some actual seismic data, and a third component that describes how close a simulated tracer test on the grid comes to observed tracer test results.

## **ADVANTAGES AND DISADVANTAGES**

### **Stochastic models as art**

If a stochastic model is being used to visually catalyze more critical thinking, then one of the primary criteria in choosing a method has to be the visual appeal of the final result. For stochastic modelling of geological objects, like sand channels, pixel-based techniques, like the sequential procedures discussed earlier, often tend to produce geological objects that are too broken up around the edges. Simulated annealing is able to better preserve the sharp edged features that many geologists expect to see. There are also object-based techniques, also often referred to as "boolean" methods, whose visual appeal is hard to beat because they deal with entire objects rather than elementary pixels.

For stochastic modelling of continuous variables, such as porosities or contaminant concentrations, a boolean approach is not going to completely do the trick. Such an approach is usually useful as a first step, but the artwork is unconvincing if continuous properties within each population are assigned randomly. The images look more plausible when the properties within

a given population have a prescribed pattern of spatial continuity. With most of the techniques for modelling continuous variables, the spatial continuity is described through a variogram model. Though experimental variograms typically show a nugget effect, experience has shown that stochastic models of continuous properties have more visual appeal if the variogram model used in the stochastic simulation procedure has no nugget effect.

For stochastic modelling of a very continuous property, such as the thickness of an aquifer, sequential techniques tend not to work well. The same variogram models that describe very continuous surfaces — the gaussian variogram model, for example, or a power model with an exponent close to 2 — also produce chaotic weighting schemes when some of the nearby samples fall directly behind closer samples. Since sequential procedures include each simulated value as hard data that must be taken into account when future nodes are simulated, and since they are typically performed on a regular grid, chaotic weighting schemes are virtually guaranteed whenever we are trying to simulate a very continuous surface.

Most of the stochastic modelling techniques for continuous variables give us the ability to honor a variogram, or a set of indicator variograms. If we are trying to produce stochastic artwork that shows specific sedimentary features, such as a succession of fining-upwards sequences, or a fabric of cross-cutting features, neither of which is well-described by a variogram, then annealing is often the most useful approach. As long as we can find a way of numerically summarizing the sedimentary features that we would like to see, then we can include appropriate terms in our energy function for annealing.

### **Stochastic models for assessing the impact of uncertainty**

When the goal of a study is to produce a handful of models that capture a couple of extreme scenarios as well as some intermediate ones, one of the practical issues that comes to the forefront is computational speed. This is especially true because it is usually not possible to steer a stochastic model to an optimistic realization or to a pessimistic one; the only way of finding such extreme scenarios is to produce many realizations and sift through them all.

It is difficult to get an objective ranking of the speed of stochastic modelling algorithms since their authors are tempted to brag that their own method is lightning fast. Though all methods are workable in practice, some require several days of run-time on fast computers to produce a single realization despite their author's enthusiastic claims of speed. Such procedures cannot be a practical basis for producing many realizations.

As a general rule, sequential procedures are slow because they have to search through an ever-increasing set of hard data to find the nearby samples that are relevant for the estimation of the local conditional probability distribution.

The notable exceptions to this are the sequential simulation procedures based on Markov random fields. These sequential procedures considerably accelerate the search for previously simulated data by using only the closest ring of data; some of them accomplish even greater speed by precalculating the kriging weights that will eventually be needed during the sequential simulation.

Annealing is a highly variable procedure in terms of its run-time. For certain problems, annealing is very rapid; for others it is abysmally slow. The key to rapid annealing is the ability to update rather than recalculate the energy function. If a global statistic has to be recalculated from scratch every time two pixels are swapped, then annealing is not likely to be speedy; if there is some convenient trick to updating the previous calculation rather than recalculating from scratch, then annealing can be very efficient.

### **Stochastic models for Monte Carlo risk analysis**

For studies that involve Monte Carlo risk analysis, there are two criteria that need to be considered when selecting a stochastic modelling technique. The first is computational speed; the second is the issue of whether the stochastic method is fairly sampling the space of uncertainty. The issue of the space of uncertainty is one of the thorniest theoretical problems for stochastic modelling. There are some who take the view that the space of uncertainty must be theoretically defined outside the context of the algorithm itself so that the user does have some hope of checking that a family of realizations represents a fair sampling of the space of uncertainty. Others, however, take the view that the space of uncertainty can *only* be defined through the algorithm and that it is, by definition, the family of realizations that is produced when all possible random number seeds are fed to the algorithm.

This issue of the equiprobability of the realizations and the fair sampling of the space of uncertainty comes to the forefront when annealing is used as a post-processor of the results from some other technique. Though there is some theory that backs up the claim that annealing fairly samples the space of uncertainty, this theory is entirely based on the assumption that the starting grid is random (or that the initial energy is very high). Though annealing offers tremendous flexibility in incorporating and honoring widely different types of information, there remains an awkward ambiguity about its implicit space of uncertainty and whether this space is being fairly sampled.

### **Stochastic models for honoring spatial variation**

If the reason for producing a stochastic model is to respect critical heterogeneities, then the choice of a stochastic modelling method should take into account the type of data that exists on heterogeneities. Annealing is a powerful technique for accommodating information on heterogeneities that comes

through complex, indirect or time-varying measurements, such as pressure transient tests in aquifers. Annealing has been used, for example, to directly honor tracer test results; though the procedure is very slow (because the tracer test response cannot be readily updated when pixels are swapped), the resulting scenarios are tremendously realistic since they capture the connectivity of the extreme values that is implicit in the actual tracer test observations.

## CONCLUSIONS

Though this paper has tended to discuss stochastic simulation methods in isolation, many practical problems call for a combination of appropriate simulation algorithms. It is unfortunate that much of the technical literature and many of the commercial software packages for stochastic spatial modelling promote a single method as better than all others. There is no stochastic modelling method that is universally best for all possible spatial phenomena. As stochastic modelling becomes more accepted in the earth sciences and as new stochastic modelling techniques are developed, it is certain that the most successful applications will be those that view the growing assortment of methods as a complementary toolkit rather than as a set of competing methods.

## REFERENCES

- Deutsch, C.V., and Journel, A.G., 1992, *GSLIB: Geostatistical Software Library and User's Guide*, Oxford University Press, 340 p.
- Yarus, J., and Chambers, R., (eds.), 1994, *Stochastic Modeling and Geostatistics, Practical Applications and Case Histories*, American Association of Petroleum Geologists.

## BIOGRAPHICAL SKETCH

R. Mohan Srivastava, is a geostatistical consultant with fifteen years of experience in the practice of geostatistics. He is an author of *An Introduction to Applied Geostatistics*, the major introductory textbook on the practice of geostatistics, and of more than thirty technical articles and reports on the theory and practice of geostatistics. In addition to using geostatistics for both exploration and development projects with the oil and mining industries, he has also applied geostatistics to a wide variety of environmental and agricultural problems. He is one of the founding partners of FSS International, a consulting group that specializes in geostatistics, and currently manages their Canadian operations from his office in Vancouver.

# Choosing and Using Simulation Algorithms

Donald E. Myers<sup>1</sup>

---

**Abstract.**--In geostatistics simulation means generating another "reality", i.e., another realization of a random function. Usually the random function is only characterized in terms of the first two moments, the mean and a spatial correlation function such as the covariance or variogram. Interpolation methods allow predicting or estimating values at non data points of the (same?) realization. Simulation algorithms generate additional realizations, often conditioned to the original data.

Originally simulation was primarily of interest as a planning tool but now some are using simulation in lieu of interpolation. Interpolation methods smooth the data, but simulation algorithms enhance the variation. In some applications the smoothing produces results that are thought to be less realistic than simulation.

There are multiple simulation algorithms available, the user must choose between them. Simulated results are often used as an input for other modeling steps. There are both practical and theoretical aspects to the problem of choosing and using a simulation algorithm. Multiple simulation algorithms will be briefly reviewed and properties discussed along with typical applications. Practical and theoretical aspects of equivalence will be discussed including the sense in which properties of the random function are reproduced

---

## INTRODUCTION

As the term is commonly used in statistics and probability, simulation means generating a set of numbers using a given probability distribution, standard limit theorems ensure that the empirical distribution (and its characteristics) will approximate those of the original distribution. This easily extends to vector valued random variables and a joint distribution. In either case the distribution function is an essential tool. The other essential tool is a random number generator for the Uniform Distribution on an interval.

In geostatistics the underlying model is a random function defined in 1,2 or 3 dimensional space,i.e., a random variable whose values are functions rather than

---

<sup>1</sup>*Professor of Mathematics, University of Arizona, Tucson, AZ 85721*

numbers. Simulation means generating an alternative "reality", i.e., another one of the functions hence preserving certain characteristics of the random function. Since simulation is a numerical technique the realization is only known at a finite number of points. In practice it is more like the problem of simulating a vector valued random variable except that the joint distribution function is seldom known. This has led to the development of a number of algorithms with differing theoretical and practical characteristics. A user must choose between algorithms and the choice of an algorithm is driven in part by the ultimate objective, i.e., the intended use of the simulated values and over time these have changed.

## **OBJECTIVES AND MODELS**

Early applications of geostatistics were principally in mining and hydrology, objectives for simulation were mainly related to these. In mining simulation was proposed as a planning tool, Journel (1974). In hydrology simulated values of hydrological parameters such as hydraulic conductivity are used as an input for flow models. Simulation can be used to design sampling plans by generating multiple realizations, optimizing the plan for each and then optimizing between plans. Because interpolation smoothes the data, interpolated surfaces are sometimes thought to be unrealistic and simulation is used instead, Dalle-Gupta et al (1995). Simulated data has also been used to study the behavior of various variogram estimators. Consider a random function model

$$Z(x) = m(x) + Y(x) \quad [1]$$

where  $Y(x)$  is a zero mean random function with some form of stationarity assumed, i.e., second order or intrinsic of order  $k$ , and  $m(x)$  is the expected value of  $Z(x)$ , usually assumed to be a linear combination of known basis functions such as monomials in the position coordinates.  $Y(x)$  is characterized by a spatial structure function such as a covariance, variogram or generalized covariance. In addition  $Y(x)$  is usually assumed to have a known univariate stationary distribution.

## **A QUICK OVERVIEW OF COMMON ALGORITHMS**

### **Turning Bands**

Let  $X(t)$  be a random function defined in 1 space,  $s$  a random unit vector in  $n$ -space and  $x$  a point in  $n$ -space then the scalar product  $\langle x, s \rangle$  is the projection of the point  $x$  onto the direction  $s$ . By integrating with respect to the random direction  $s$ , one obtains a random function  $Z(x)$  defined in  $n$ -space. This provides a relationship between the covariance function  $C_1(r)$  for  $X(t)$  and the covariance  $C_n(r)$  of  $Z(x)$ . Given a covariance for  $n=3$  one can easily solve for the corresponding covariance in 1-space. Then by writing the 1-dimensional covariance as a convolution one

obtains a function that can be used in a Box-Jenkins, i.e., moving average, simulation algorithm (in 1-space). To discretize the algorithm the user must make choices, first, one can not use an infinite number of directions. They should be equally spaced and in 3-dimensions geometry intervenes to show that 15 is maximum number. Theoretically the Moving Average simulation is an integral and must be approximated by an infinite series but only a finite number of terms are used. Instead of simulating a value at each point on a line, one simulates a value for each of a set of disjoint intervals. When the lines are rotated in different directions we obtain the "turning bands". Given a point  $x$  then one determines the interval on each direction that the point projects into and looks up the simulated value for that interval. The simulated value for the point  $x$  is then simply a sum of those values.

While the integral equation linking the 1-dimensional covariance and the  $n$ -dimensional covariance is easily solved for  $n=3$ , it is not so simple for  $n=2$ . This led to the use of spectral methods, see Mantoglou and Wilson (1982) for applications in hydrology and Booker (1985) for solutions of the integral equation when  $n=2$ . There is another important difference between the cases for  $n=2$  and  $n=3$ . For any even integer it is easy to choose equally spaced directions in 2-space (equally spaced points on the unit circle). There is no upper bound on the number of equally spaced directions.

The Turning Bands algorithm will reproduce the covariance function in an "average" sense, i.e., if one generates multiple realizations then averages the sample covariance functions (or sample variograms) the average will approximate the original covariance (variogram).

### **L-U Decomposition**

Given locations  $x_1, \dots, x_m$  the covariance matrix of the random variables  $Z(x_1), \dots, Z(x_m)$  is positive definite and may be "factored" in at least two ways. The Cholesky decomposition produces two matrices  $L$ ,  $U$  (which are respectively lower triangular and upper triangular) such that  $L^T = U$ . Now suppose that  $Y_1, \dots, Y_m$  are uncorrelated, mean zero, variance one random variables. The product  $UY$  is a collection of random variables whose covariance matrix is the original. Although the technique was well-known in statistics, the application and extension to spatial data is found in Davis (1987) where it is also shown that the conditioning is easily incorporated. There are at least two difficulties; most algorithms for finding the L-U decomposition require that the entire matrix be in memory at one time and this limits the size of the grid to about 1000 points. Alabert (1987) has given a modified form of the algorithm allowing the use of a moving neighborhood for the conditioning and hence the use of multiple covariance matrices instead of one large matrix. Since the matrix "slides" over the grid one must still consider the possible discrepancies that might be generated by having a grid location occur in multiple

matrices.

Although it does not completely solve the problem, Davis (1987) also proposed the use of the square root of the covariance matrix which is then approximated by a polynomial expansion which only requires computing powers of the matrix also see Dietrich and Newsam (1995). Myers (1989) extended both of these algorithms to the vector case including the undersampled case for conditioning.

One advantage of the L-U or square root algorithms is that they are dimension independent but both suffer the same difficulty with respect to the univariate distribution as does the Turning Bands method, namely that the simulated values are obtained as linear combinations of uncorrelated random variables. Hence a transformation to normality is needed along with a reverse transformation. There is a parallel processing version of the Cholesky algorithm which would allow the use of larger covariance matrices, i.e., larger grids, but it seems not to have been used in geostatistical applications as yet.

### **Sequential Gaussian**

It is well-known that in the case of multivariate normality the Simple Kriging estimator is the conditional mean, i.e., the mean of the conditional distribution, and the Kriging variance is the conditional variance. The Sequential Gaussian simulation algorithm is based on this property. Given a set of data locations and a grid point to be simulated, one computes the Simple Kriging estimated value for the grid point and the kriging variance. These are used in the conditional normal distribution which is then "sampled". Having generated a value at the grid point, this location and value are considered to be a part of the data set and process is repeated at a new grid point. Since it is straightforward to use a moving search neighborhood for the simple kriging it is not necessary to use the entire "data" set for each grid point but rather only those that are close. The user must then select the parameters for the search neighborhood and also select the order in which grid points are simulated since points selected earlier may be used in the subsequent simulations.

Since normality is explicitly used in the algorithm it is then necessary to transform the data and to re-transform the simulated values. As with the Turning Bands and L-U decomposition methods, the covariance used pertains to and is modeled from the transformed data. In an average sense it can be shown that any randomly selected sequence of grid points produces the same set of realizations.

### **Simulated Annealing**

The title is somewhat mis-leading since Simulated Annealing is really an optimization method. A realization is generated by minimizing the "difference"

between the sample variogram (or sample covariance) and the theoretical model. First the data is used to model a probability distribution, then using this distribution and a random number generator, a value is simulated at each grid location (independently). Although this will ensure that the univariate distribution is preserved it would in general result in a pure nugget effect variogram. Next select at random two locations, interchange the values at these two points and compute the sample variogram. If the new sample variogram is closer the interchange is retained, otherwise not. New locations are selected and the process repeated. A "temperature" function is used to ensure that the iteration stops. For a brief introduction see Deutsch(1995), Deutsch and Cockerham (1994).

### VARIATIONS AND EXTENSIONS

The four methods described above essentially assume that the random function is continuous valued. At least two variations occur on this theme, coded data obtained by for example an indicator transform applied to continuous data, and categorical data. One might simulate indicator transform data for reasons that are essentially the same as those related to the use of indicator kriging as opposed to kriging of the original variable. Journel and Isaaks (1984) use an indicator transform to separate the data corresponding to a high grade mineralization from that corresponding to a low grade mineralization. Bierkens and Weerts (1994) apply indicator coding to categorical data, e.g., soil texture classes. In this case it is necessary to assign an (arbitrary) ordering to the texture classes. The use of indicators essentially corresponds to assigning a probability distribution to each point, i.e., one is simulating a probability distribution at each point. This distribution is conditioned on the "data". As in indicator or probability kriging, if the values are kriged or simulated for each cut-off separately then the order relations may not be satisfied and adjustments are needed.

### SOME QUESTIONS

Usually one desires that the univariate marginal distribution be preserved and this also creates a problem. First of all one must distinguish between a theoretical distribution for the random function model and the spatial frequency distribution of a particular realization although the latter is what is usually computed. Since the simulated value at any point on the grid is obtained as a sum of independent, i.e., uncorrelated, random numbers one must expect a form of the Central Limit Theorem effect. Actually one need not resort to the Central Limit Theorem, there are only a few distribution types for which the sum of independent random variables has the same distribution type as the summands. These include the Normal and the Gamma, hence it is common to utilize a transformation to transform the original data frequency distribution to Normal and then to re-transform the simulated values back. Since the original data set is finite, the empirical distribution will be discrete and

hence some smoothing is needed to transform to Normal, the transformation is not uniquely determined but requires some user choices. In addition the covariance used in the simulation step reflects the spatial structure of the transformed data and in general there is no assurance that the spatial structure of the original data is preserved in the process. A desire to simplify the algorithms gave rise to an emphasis on the LU decomposition and sequential gaussian methods both of which use normality more explicitly. As the need for application to non-continuous variables and also for simulation on much larger grids increased, other algorithms were introduced. The choice between the algorithms was and is largely driven by personal preference or practical aspects of the application. Underlying this diversity of algorithms was an implicit but never stated assumption that there was some form of equivalence and hence the difference was only computational. This might be interpreted as meaning that if one considers the set of all possible realizations that could be generated (for a given set of parameter assumptions) was the same for all methods or at least approximately so. Since in practice one only generates a small number of realizations, i.e., the values at a finite number of points for each of a small number of realizations there is another implicit assumption, that for the particular purpose or objective the conclusions were essentially the same irrespective of which small set of realizations is used. Neither of these implicit assumptions has really been tested or even considered, most users do not use multiple algorithms and make comparisons nor do they generate multiple finite sets of realizations to compare between the sets.

Characterizing or defining "equivalence" is non-trivial. One very strong perception would be that if one considers the totality of possible realizations from one algorithms vs another then the two are equivalent if the two sets are the same. Unfortunately the algorithms in general only depend on second moment properties and hence do not uniquely determine a random function. One might also consider equivalence in a numerical sense, two algorithms are equivalent if for any finite set of grid points, the same sets of values are generated by both algorithms. While equivalence is principally a theoretical question, if two algorithms are not equivalent then one must ask how to decide between algorithms and to what extent is the validity of the conclusions related to the choice of the algorithm.

All of the simulation techniques depend in a critical way on the use of a random number generators. It is well-known that such numbers are in fact not really random but by knowing the "seed" the sequence is completely predictable. However there are more serious problems, there are many different random generator algorithms and they are not equivalent nor do they have the same characteristics. Some are highly dependent on hardware and all depend on the word length in memory. A highly touted "portable" random number generator has recently been shown not to be machine independent as earlier believed. As yet there has been little study of the effect of different random number generators on the various simulation algorithms.

## REFERENCES

- Alabert, F. 1987. The practice of fast conditional simulations through the LU decomposition of the covariance matrix. *Math. Geology* 19, 369-386
- Bierkens, M.A.P., and Weerts, H.J.T., 1994, Application of indicator simulation to modelling the lithological properties of a complex confining layer. *Geoderma* 62, 265-284
- Booker, P.I. 1985. Two dimensional simulation by Turning Bands. *Math. Geology* 17, 81-90
- Dalle-Gupta, A., Lake, L. and Pope, G.A., 1995, Characterizing heterogeneous permeable media with spatial statistics and tracer data using sequential simulated annealing. *Math. Geology* 27, 763-788
- Davis, M.W. 1987. Production of conditional simulations via the LU triangular decomposition of the covariance matrix. *Math. Geology* 19, 91-98
- Davis, M.W. 1987. Generating large stochastic simulations-The matrix polynomial approximation approach. *Math. Geology* 19, 99-108
- Dietrich, C.R. and Newsam, G.N., 1995, Efficient generation of conditional simulations by Chebychev matrix polynomial approximations to the symmetric square root of the covariance matrix. *Math. Geology* 27, 207-228
- Dimitrakopoulos, R. 1990. Conditional simulation of intrinsic random functions of order k. *Math. Geology* 361-380
- Deutsch, C.V., 1995, Introduction to the application of simulated annealing for geostatistical applications. *GEOSTATISTICS Newsletter* 7, 8-14
- Deutsch, C.V. and Cockerham, P.W., 1994, Practical considerations in the application of simulated annealing to stochastic simulation. *Math. Geology* 26, 67-82
- Easley, D. H., Borgman, L.E. and Weber, D., 1991, Monitoring well placement using conditional simulation of hydraulic head. *Math. Geology* 23, 1059-1080
- Gutjahr, A., Bullard, B. and Hatch, S., Joint conditional simulations and flow modeling. in *Geostatistics for the Next Century*, R. Dimitrakopoulos (ed), Kluwer Academic Pub. Dordrecht 185-196
- Journel, A.G. 1974 Geostatistics for conditional simulation of ore bodies. *Economic Geology* 69
- Journel, A.G. and Isaaks, E.H. 1984. Conditional indicator simulation: Application to a Saskatchewan uranium deposit. *Math Geology* 16, 685-718
- Mantoglou, A. and Wilson, J.L. 1982. The turning bands method for simulation of random fields using line generation by a spectral method. *Water Resources Res.* 18, 1379-1384
- Myers, D.E., 1989, Vector conditional simulation. in *Geostatistics*, M. Armstrong (ed), Kluwer Academic Publishers, Dordrecht, 283-292

# The Components of Geostatistical Simulation

Carol A. Gotway<sup>1</sup> and Brian M. Rutherford<sup>2</sup>

---

**Abstract.** There are many approaches to geostatistical simulation that can be used to generate realizations of random fields. These approaches differ fundamentally in a number of ways. First, each approach is inherently different and will produce fields with different statistical and geostatistical properties. Second, the approaches differ with respect to the choice of the features of the region that are to be modeled, and how closely the generated realizations reproduce these features. Some fluctuation in the statistical and geostatistical properties of different realizations of the same random field is natural and desirable, but the proper amount of deviation is an open question. Finally the approaches differ in how the conditioning information is incorporated. Depending on the source of randomness and the uncertainty in the given data, direct conditioning of realizations is not always warranted. In this paper, we discuss and illustrate these differences in order to emphasize the importance of these components in geostatistical simulation.

---

## INTRODUCTION

Geostatistical simulation, as illustrated in Figure 1, provides a way to quantify uncertainty in the prediction of a complex system response. Some information is usually available on a process of interest (e.g., measurements on a hydrogeologic parameter of interest in a particular region), but the transfer function (e.g., a groundwater flow model) may require a detailed spatial map of the parameter on which the measurements were obtained. (Such a map can be created through spatial prediction techniques such as kriging.) The value of the transfer function can then be computed over the mapped values to obtain a prediction of a system response (e.g., groundwater travel time), but an estimate of the uncertainty associated with the prediction, often as important as the predicted value itself, is typically much more difficult to infer. One alternative is to generate realizations of the random field that share the available information on the parameter of interest and use the transfer function to compute a system response for each. If the realizations characterize the spatial uncertainty in the parameters of interest, the resulting distribution of predicted system responses will reflect this uncertainty.

---

<sup>1</sup> Assistant Professor of Biometry, University of Nebraska, Lincoln, NE

<sup>2</sup> Senior Member of Technical Staff, Sandia National Laboratories, Albuquerque, NM

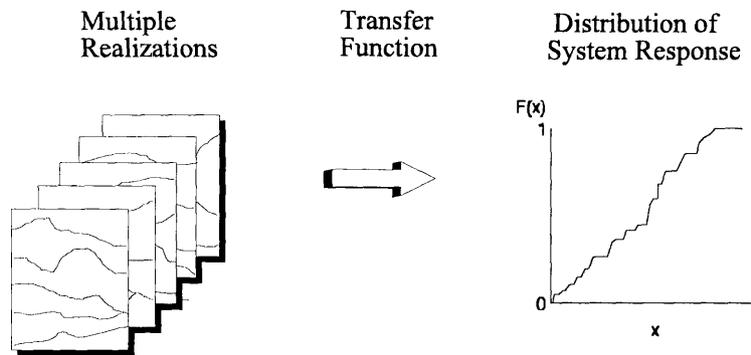


Figure 1. Illustration of Stochastic Simulation (Journel, 1989).

Although originally introduced as a geostatistical approach to uncertainty analysis (Journel, 1989), geostatistical simulation has grown to provide a variety of new approaches to stochastic modeling that can incorporate complex spatial relationships. These methods are frequently used in hydrology, petroleum engineering, and the environmental sciences, not only for uncertainty analysis, but also for risk analysis, decision analysis, and stochastic modeling.

Many different approaches to geostatistical simulation can be used to construct the realizations. For continuous variables these include the turning bands method (Journel, 1974), spectral methods (Borgman et al., 1984; Gutjahr, 1989), LU or Cholesky decomposition (Alabert, 1987; Davis, 1987), sequential Gaussian simulation (Journel, 1989), and fractal approaches (Hewett and Behrens, 1990). Approaches to simulation of discrete or categorical variables include Boolean methods (Serra, 1982; Jeulin, 1987; Chautru, 1989) and the more general random closed set approaches (Stoyen et al., 1987), and the use of truncated Gaussian random functions (Matheron et al., 1987). Sequential indicator simulation (Journel and Alabert, 1989), simulated annealing (Kirkpatrick et al., 1983; Deutsch and Journel, 1992) and genetic algorithms (Whitley, 1994) can be used with both continuous or categorical variables.

For any particular type of data (continuous, discrete, or categorical), different approaches will assign different statistical and spatial features to the generated fields. These differences could be due to: (1) differences in the statistical model used to characterize the phenomenon under study. These differences are most obviously represented by the criteria used to describe spatial data relationships and characterize

available information. We refer to such criteria as *summary functions*. Examples of summary functions include the univariate data distribution, the covariance function, indicator semivariograms, and the shape and intensity of primary elements in Boolean simulations; (2) differences in algorithms (e.g., using sequential Gaussian simulation versus using LU decomposition); (3) the acceptable deviation between the summary functions specified to the algorithm and those calculated from resulting realizations; and (4) the use of conditioning information. Since all of these differences will translate into differences in the nature of the spatial heterogeneity reflected in the resulting realizations, and hence into differences in uncertainty assessments, risk calculations, and the conclusions drawn from them, it is important to consider each of these components carefully when using geostatistical simulation. In the remainder of this paper, we elaborate on these components, discuss some of the different choices within each component, and illustrate the effect of these differences on the nature of the simulated fields and on functions derived from them.

## **COMPONENTS OF GEOSTATISTICAL SIMULATION**

When using geostatistical simulation to characterize the spatial variation of parameters of interest, the modeler has to make several important choices. Each choice will affect the nature of the resulting realizations and the values of functions derived from them. Thus, careful thought with regard to the major components of geostatistical simulation is crucial in modeling the uncertainty of spatial functions.

### **Choice of statistical model and selection of summary functions**

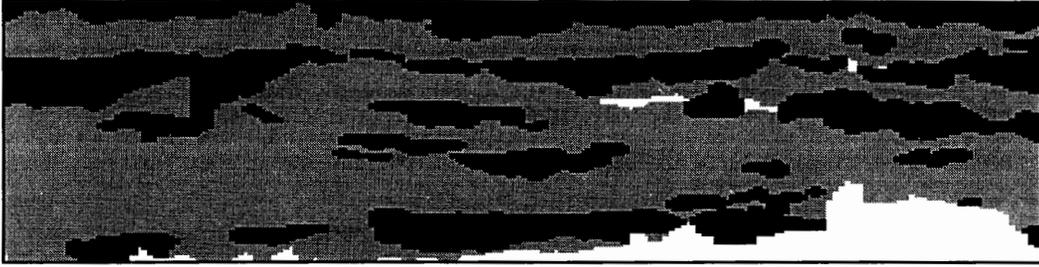
For geostatistical simulation of a spatial process, a statistical model must describe a multivariate probability distribution for the spatial data of interest. Given such a distribution, realizations can be generated by sampling randomly from this specified distribution. For example, in LU decomposition, the underlying multivariate distribution is Gaussian, and the parametric form for this distribution can be found in most introductory statistical texts. Similarly, general formula for the multivariate distributions of Boolean random sets, as well as formula for distributions of other random sets, are given in Stoyen et al., (1987). In other cases, such as the sequential indicator simulation model, or certain Markov Random Field Models, the exact form of the multivariate distribution is intractable or unknown, and is conditionally specified using a local neighborhood structure.

Once a particular model has been chosen, the multivariate distribution is specified through the use of summary functions. For example, generation of realizations from a multivariate Gaussian distribution requires specification of the process mean and

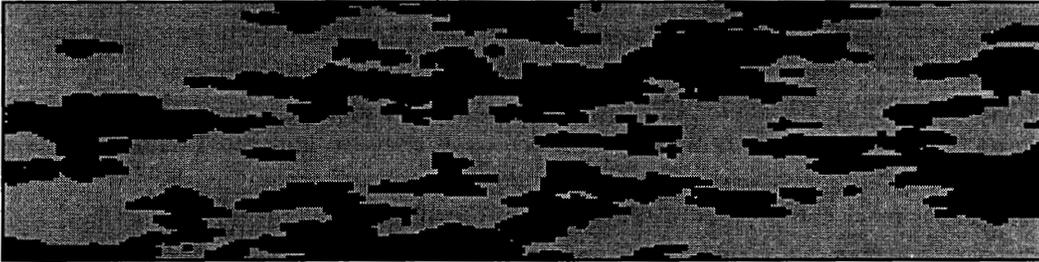
covariance structure which is typically determined through the semivariogram. In Boolean models, the multivariate distribution depends on the intensity of the Poisson process, the shape of the primary elements, and the probability distributions of their sizes. In sequential indicator simulation, the multivariate distribution is specified nonparametrically through the use of indicator semivariograms. Consequently, using a particular simulation algorithm that requires a specified set of summary functions is equivalent to specifying a multivariate probability model for the spatial process. These models necessarily emphasize different spatial features of the process, so it is important to determine what features are most important in the application of interest and how best to quantify them from the available information. In applications where a transfer function is evaluated over the generated realizations, it is important to determine the features of the process that will greatly influence the value calculated through the transfer function. For example, in reservoir modeling, it is often important to characterize those features that influence the flow of fluids through reservoirs, and in environmental risk analysis, it may be important to characterize those features that control the spread of a contaminant of special interest.

Emphasizing different features through the use of different summary functions can give rise to realizations that appear quite different. As an example, consider the top image in Figure 2. It shows a geologic section (obtained from digitization) of interbedded debris flow (fine grained) and sheet flood (coarse grained) units produced as part of site-characterization studies at the Greater Confinement Disposal Facility in Southern Nevada (Rutherford and Gotway, 1994; McCord et al., 1996). Recharge through sections of this type is an important component of such studies and the position, shape, and amount of each type of unit are important for modeling vadose zone flow. The last three images in Figure 2 are unconditional realizations generated from statistics inferred from the training image using sequential indicator simulation, Boolean simulation, and a more general random set approach. Sequential indicator simulation uses two summary functions, the proportion of debris flow and the indicator semivariogram. The indicator semivariogram attempts to incorporate the connectivity of the permeable and impermeable zones. Although the shape of the debris flow units can be controlled to some degree by the anisotropy reflected in the indicator semivariograms, the shapes are somewhat limited. The Boolean model uses summary functions that specify the intensity of the process, and the shape and size of the debris flow units. Here the emphasis is on shape, rather than connectivity. The last realization is based on the theory of random closed sets (Kendall, 1974). In this application the summary functions are described by a system of structuring elements, and any random set can be characterized through its intersection with these elements. The last realization was generated using 19 structuring elements consisting of lines, triangles and rectangles. Here there is an attempt to account for both shape and connectivity. The use of multiple-point statistics (Deutsch and Journel, 1992; Guardiano and Srivastava, 1993) might also be used to incorporate

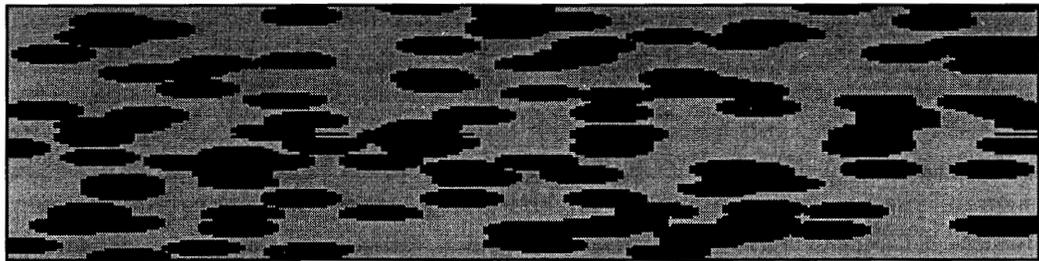
**GCD Training Image**



**Sequential Indicator Realization**



**Boolean Realization**



**Hitting Functions Realization**



Figure 2. The Use of Different Summary Functions in Lithologic Simulation.

more features of importance. Note that in such applications, the goal is not to reproduce the training image, but to capture essential features of the image (not always visible by eye) that are most important to the transfer function of interest.

### **Choice of Simulation Algorithm**

In addition to the selection of an appropriate set of summary functions that determine the modeling approach, different algorithms for generating the realizations often impact the results. In implementing most geostatistical simulation algorithms, there are a number of parameters that must be made specific, often based on a tradeoff between speed and accuracy. Examples include the number of turning bands lines, the number of conditioning values retained in sequential simulation, the choice of annealing schedule and convergence criteria, and the number of structural elements in random set generation. In many cases, blatantly improper choices can be clearly reflected in the realizations, e.g., the appearance of banding effects in turning bands method when the number of lines is small. In other cases, apparently reasonable selections can also have an effect on the realizations and on the resulting uncertainty distributions derived from them. In a comparative study of several geostatistical simulation algorithms using a variety of exhaustive data sets and transfer functions, Gotway and Rutherford (1994) illustrate these algorithmic differences. Figure 3 shows uncertainty distributions of the average of local geometric means computed from conditional realizations of a Gaussian field. Although the shape of the uncertainty distributions is similar, they differ in the upper tail which could be crucial for risk assessments. Comparing the distributions obtained using LU Decomposition, Sequential Gaussian, and Turning Bands algorithms shows the effect of algorithmic differences on the response uncertainty distributions. Comparing the distribution obtained using sequential indicator simulation to the others illustrates the effect of choice of summary functions. As shown in Gotway and Rutherford (1994), these differences are much more pronounced for unconditional realizations.

### **The Acceptable Deviation**

The acceptable deviation refers to the difference between the summary functions calculated from realizations and those specified by the model. For example, should the variogram calculated from each realization exactly match the variogram specified? Certainly, the answer to this question is no, since, in simulation, some deviation of summary functions is desirable in order to reflect uncertainty. However, the amount of deviation acceptable is an open question. Unless there is

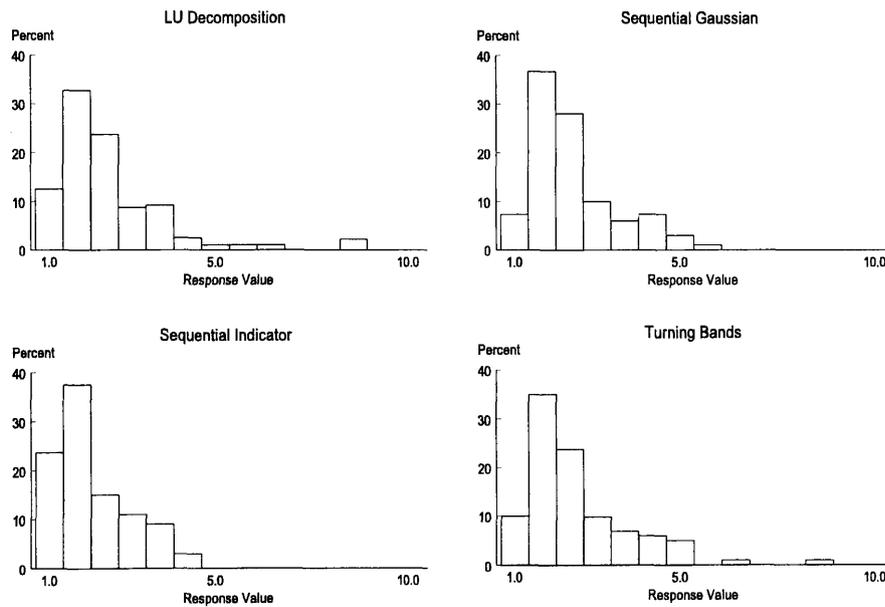


Figure 3. Response Uncertainty Distributions.

“post-generation” alteration to the realizations, these deviations depend on the specified summary functions and on the simulation algorithm. Often, simulation algorithms are compared and evaluated based on their ability to reproduce the desired functions over an average of realizations. However, the variability in the summary functions can be quite different depending on the particular algorithm used and still, on the average, reproduce the specified summary functions. More importantly, such a reproduction assumes that the summary functions specified are in fact known, when in reality they are estimated from available data.

The fundamental concern with the acceptable deviation is that differences between specified summary functions and those calculated from generated realizations should reflect: (1) the uncertainty in input summary functions; (2) the natural variability that comes with random spatial sampling over a finite region to generate realizations, and not (3) the variability (or lack of variability) introduced by the simulation algorithm. None of the available algorithms account for (1), and with the exception of algorithms in which a multivariate distribution is directly sampled (such as LU decomposition) it is also not clear whether (2) is accomplished correctly. For many applications, (3) may be the primary determinate of deviations between theoretical and reproduced summary functions.

## The Use of Conditioning Information

Conditioning a simulation on known values or soft information (incomplete information concerning the region of interest) must be done in a way that reflects the true conditions imposed on the simulated region by that data or information. At present, the common approach for algorithms using a sequential method of generation is to treat previously generated values as known data values in the simulation. The study in Hansen (1992) indicates that this approach might “over-condition” or “under-condition” the realizations. The result is a biased uncertainty distribution of response that may be unrealistically wide or too narrow. Moreover, if data are measured with error, conditioning on these data may not be warranted; rather some filter must first be applied to reduce the noise introduced by measurement error.

In many applications, the summary functions are derived from a training image, and simulation of complex geologic structures is often based on a training image that is derived from outcrop mapping (e.g., Figure 2). However, the goal of simulation is not to reproduce a training image. Rather, the objective is to generate realizations that account for information believed relevant to the general process under study. Such information might include the general shape and relative distributions of important units, connectivity patterns, cyclic trends, etc., that could be expected to be general properties of the process. It most probably does not include exact values obtained from the training image. In fact, Guardiano and Srivastava (1993) note that “the training image is itself is not the underlying phenomenon. It is assumed, however, that both share some structural properties such as specific multiple-point covariances.” Consequently, it seems very difficult to justify using the exact values (or even some randomly selected subset of them) as conditioning information.

## REFERENCES

- Alabert, F.G. 1987. The practice of fast conditional simulations through the LU decomposition of the covariance matrix. *Mathematical Geology*, 19: 369-386.
- Borgman, L.E., M. Taheri, and R. Hagen. 1984. Three-dimensional, frequency-domain simulations of geological variables. In G. Verly et al. (eds.), *Geostatistics for Natural Resources Characterization*. Reidel, Dordrecht, pp. 517-541.
- Chautru, J.M. 1989. The use of Boolean random functions in geostatistics. In M. Armstrong (ed.), *Geostatistics*, Vol. 1, Kluwer, Dordrecht, pp. 201-212.
- Davis, M.W. 1987. Generating large stochastic simulations via the LU triangular decomposition of the covariance matrix. *Mathematical Geology*, 19: 91-98.

- Deutsch, C.V. and A.G. Journel. 1992. Annealing techniques applied to the integration of geological and engineering data. Stanford Center for Reservoir Forecasting, Stanford, CA.
- Gotway, C.A., and B.M. Rutherford. 1994. Stochastic simulation for imaging spatial uncertainty: Comparison and evaluation of available algorithms. In M. Armstrong and P.A. Dowd (eds.), *Geostatistical Simulations*, Kluwer, Dordrecht, pp. 1-21.
- Guardiano, F. And M. Srivastava. 1993. Multivariate geostatistics: Beyond bivariate moments. In Soares, A. (ed.), *Geostatistics Troia '92*. Kluwer, Dordrecht, pp. 133-144.
- Gutjahr, A. 1989. *Fast Fourier Transforms for Random Fields*. Technical Report No. 4-R58-2690R, Los Alamos, NM.
- Hansen, K.M. 1992. The use of sequential indicator simulation to characterize geostatistical uncertainty. Technical Report No. SAND91-0758, Sandia National Laboratories, Albuquerque, NM.
- Hewett, T. And R. Behrens. 1990. Conditional simulation of reservoir heterogeneity with fractals. *Formation Evaluation*, 330-310.
- Jeulin, D. 1987. Anisotropic rough surface modeling by random morphological function. *Acta Stereologica*, 6:183-189.
- Journel, A.G. 1974. Geostatistics for conditional simulation of ore bodies. *Economic Geology*, 69:673-687.
- Journel, A.G. 1989. *Fundamentals of Geostatistics in Five Lessons*. American Geophysical Union, Washington, D.C.
- Journel, A.G., and Alabert, F. (1989). Non-Gaussian data expansion in the earth sciences. *Terra Nova*, 1:123-124.
- Kendall, D.G. 1974. Foundations of a theory of random sets. In E. Harding and D. Kendall (eds.), *Stochastic Geometry*, Wiley, Chichester, pp. 322-376.
- Kirkpatrick, S., C.D., Gelatt, and M.P. Vecchi. 1983. Optimization by simulated annealing. *Science*, 220:671-680.
- Matheron, G., H. Beucher, C. De Fouquet, A. Galli, D. Guerillot, and C. Ravenne. 1987. Conditional simulation of the geometry of fluvio-deltaic reservoirs. SPE Paper No. 16753, 62nd Annual Technical Conference and Exhibition of the Society of Petroleum Engineers, Dallas, TX., September 27-30.
- McCord, J.T., C.A. Gotway, and S.H. Conrad. 1996. Inferring recharge rates using environmental tracers in heterogeneous vadose zones. *Water Resources Research*, to appear.
- Rutherford, B.M., and C.A. Gotway. 1994. Selecting features from spatial data for use in stochastic simulation. *Proceedings of the 35th Annual Meeting of the Institute of Nuclear Materials Management*, July 17-20, Naples, FL.
- Serra, J. 1982. *Image Analysis and Mathematical Morphology*. Academic Press, London.
- Stoyen, D., W.S. Kendall, and J. Mecke. 1987. *Stochastic Geometry and Its Applications*. Wiley, New York.
- Whitley, D. 1994. A genetic algorithm tutorial. *Statistics and Computing*, 4: 65-85.

# A General Method for the Efficient Selection of Sampling Locations for Problems in Environmental Restoration

Brian M. Rutherford<sup>1</sup>

---

**Abstract.**—Problems in environmental restoration that involve detecting or monitoring contamination or site characterization often benefit from procedures that help select sampling or drilling locations for obtaining meaningful data that support the analysis. One example of this type of procedure is a spatial sampling program that will "automatically" (based on the implementation of a computer algorithm) guide an iterative investigation through the process of site characterization at a minimal cost to determine appropriate remediation activities. In order to be effective, such a procedure should translate site and modeling uncertainties into terms that facilitate comparison with regulations and should also provide a methodology that will lead to an efficient sampling plan over the course of the analysis. In this paper, a general framework is given that can accomplish these objectives and can be applied to a wide range of environmental restoration applications. The methodology is illustrated using an example where soil samples support the characterization of a chemical waste landfill area.

---

## INTRODUCTION

The objectives of environmental restoration projects often require data for detecting, monitoring, or estimating the extent of contamination. Spatial sampling procedures may be required to help locate soil sample or bore hole locations that will provide the most additional site information at the lowest cost. Two examples of problems that might benefit from this type of procedure follow:

a) Find drilling locations for a monitoring system that will provide the information required to minimize the probability that a contaminant will cross a site boundary undetected; and

b) Select samples to reduce as much as possible the area or volume of a 90% confidence region for a specific concentration contour.

---

<sup>1</sup> Senior Member Technical Staff, Sandia National Laboratories, Albuquerque, NM.

Two essential elements of procedures designed to accomplish these types of environmental restoration project objectives are: I) to translate site and modeling uncertainties into terms that permit comparison with regulations so that compliance and cleanup requirements can be assessed; and II) to provide methodology that can lead to an efficient sample plan throughout the course of the analyses. The concepts involved with these two elements of the procedure provide the basis for the remainder of this paper.

### TRANSLATING SITE AND MODELING UNCERTAINTIES INTO TERMS OF REGULATORY COMPLIANCE

Limited sampling and other factors contribute to site and modeling uncertainty. Most site characterization applications include uncertainty from a number of sources: contaminant values at locations that have yet to be sampled are unknown; flow and geostatistical models have uncertainty associated with them; and data available from samples collected at the site and "soft data" relevant to the site, collected through a number of alternative characterization methodologies have varying degrees of uncertainty associated with them.

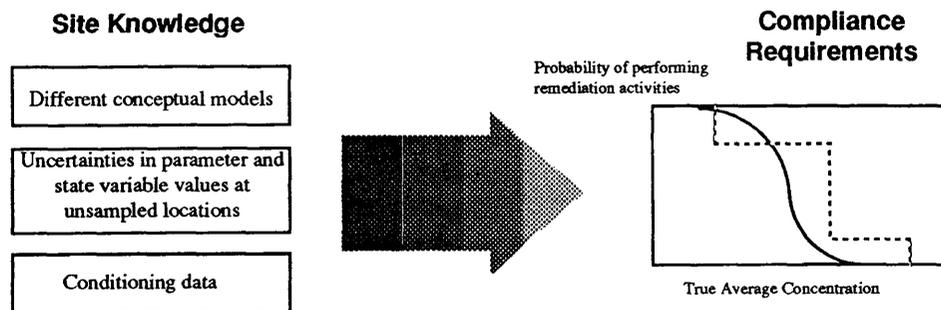


Figure 1

Recognizing these uncertainties as unavoidable, environmental regulations should be and often are stated probabilistically. Constraints on the allowable probability of exceeding a regulatory threshold, stated as a function of the (hypothetical unknown) true level of contamination, provides one example. In order to be effective, the procedure must translate site uncertainties into the probabilistic framework of these regulations. Figure 1 illustrates this element of the process for the (simpler than most) problem of characterizing average contamination levels. This will provide those responsible for demonstrating compliance of a site, or those responsible for analyzing cleanup alternatives and capabilities, with a way of making critical decisions.

The examples discussed earlier will be different (and generally more complex). Example a), for example, might have a vertical scale indicating the detection probability. The computation of these probabilities would depend on the assumed flow models and their parameters as well as the geological features of the site.

This translation of site uncertainties into a regulatory framework can be achieved through a two-step process. The first step is to convert site and modeling uncertainties into simulated realizations, each realization providing a possible map of site contamination. When taken collectively, the realizations reflect probabilistically the way that contaminant levels at the site might be configured. These realizations are then evaluated according to an objective function that calculates the compliance probability or risk.

The remainder of this section is partitioned into subsections that address these three components of this process

### **Factors Contributing to Site and Modeling Uncertainty**

At least five general sources might contribute to uncertainty in the estimated concentration levels of contaminants at a site:

- 1) Uncertainties in small scale hydrogeologic property values, either in parameter values or in state variables. These values are uncertain at locations that have not been sampled;
- 2) Uncertainty introduced by incomplete knowledge regarding the location and extent of large scale geological features such as stratigraphic units or zones of heavy fracturing;
- 3) Uncertainty concerning the extent and configuration of the contaminant source;
- 4) Uncertainty concerning the appropriate flow and transport models and in parameters associated with these models; and
- 5) Uncertainty in analytical results for samples already collected and analyzed.

The first step of the translation process is to convert site and modeling uncertainties into simulated realizations that differ from one another to an extent that reflects these uncertainties. Each source listed above must be taken into

account in the simulation process to reflect its impact on differences between simulated realizations.

Fields of hydrogeologic parameter values can be generated using stochastic simulation techniques conditioned on sample data where available. Replication of this simulation process can account for parameter uncertainty source (1). Uncertainties in state variables are often estimated through repeated application of the flow models using different sets of the hydrogeological and other parameter values (selected through simulation) for each replication and conditioned based on known values of state variables where applicable. Other parameter values may include the specification of large scale geological features, source (2) above, or contaminant source, strength and configuration, source (3). Probabilities for different contaminant sources and for different scenarios involving large scale geological features must be assessed (perhaps subjectively) and introduced into the simulation process accordingly.

Modeling uncertainty, source (4), can be accounted for by selecting among alternative models with probability of selection consistent with each model's likelihood of being applicable (again, likely to be a subjective decision). In most cases, the appropriate models are established, but parameters associated with these models must be specified. Their uncertainty distribution must be estimated, and the parameters generated randomly through Monte Carlo or perhaps alternative simulation techniques. The fifth source of uncertainty -- that resulting from inexact analytical procedures, can be accounted for by reducing the conditioning requirements of the simulation and consequently allowing differences between values obtained through sampling and those obtained through simulation at the same location.

Together, accommodating for these sources of uncertainty through the simulation process will, if performed correctly, provide an estimated probability space for site reality that is both unbiased and provides an accurate assessment of uncertainty based on present site knowledge. This discrete (by necessity) representation of the probability space is described next.

### **The Simulated Realizations**

The simulated realizations represent an intermediate step in the translation process. Each realization provides one possible map of site contaminant levels. At each grid location, the set of realizations yields a histogram of possible values. Figure 2 illustrates how the set of realizations can be interpreted in this way.

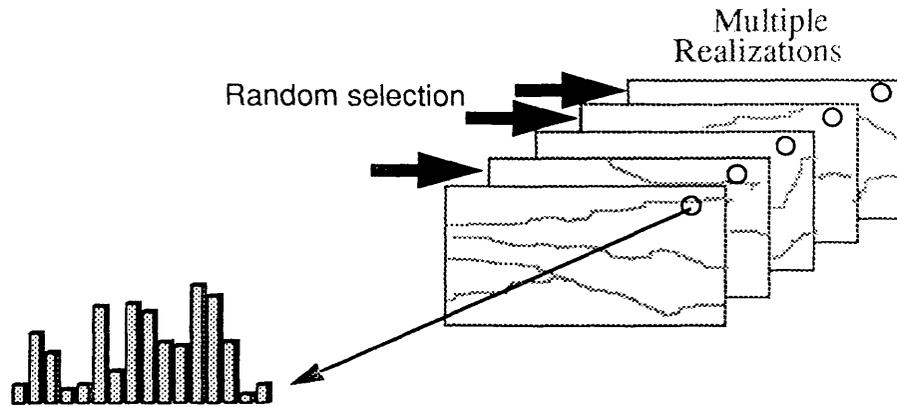


Figure 2

An appropriate interpretation of this set of realizations, analyzed pointwise, is that the probability of a contaminant value at this location, smaller than any level  $x$ , can be estimated by the ratio of the number of realizations that have a value less than  $x$  at this location, to the total number of realizations. This probabilistic interpretation can be extended to multiple locations or to the entire region. The probability that no point in the entire region exceeds the level  $y$ , for example, can be estimated by the ratio of the number of realizations where  $y$  is not exceeded at any point in the region to the total number of realizations.

This interpretation of the set of realizations is utilized in two ways in the methodology presented in this report. It provides justification for the probabilistic interpretation of the values indicating likelihood of compliance or risk functions that are compared to regulations and it provides a basis for the resampling procedure discussed in the subsection Determining an Efficient Sample Plan.

### The Objective Function

The objective function is a transformation that will compute a value indicating the level of compliance based on a complete specification of all site parameters and state variables. For the present problem, the objective function is used to transform the set of simulated realizations into a probabilistic expression that relates to regulatory compliance. Examples of the roles that objective functions might play in the example problems listed in the Introduction Section are:

- a) Provide an estimate of the probability that a contaminant will cross a site boundary undetected by a specific monitoring system within 10 years; and
- b) Provide a distribution showing probabilistically the spatial or volumetric extent of the 90% confidence region for a specific concentration contour.

## DETERMINING AN EFFICIENT SAMPLING PLAN

The process of sample location selection is often an iterative process. As new data are obtained, models are updated to provide more accurate estimates of site contamination levels or predictions of flow and transport and a reassessment of future data needs is performed. A second essential element of an effective spatial sampling procedure is that it utilizes data available at every sampling iteration and that models based on these data produce an efficient sample plan -- one where maximum relevant information is provided at the lowest possible cost. At each iteration, the next set of sample locations is selected from numerous potential supplemental sample sets. Methods for choosing sets to consider and methods for evaluating those that are considered must be carefully selected.

A number of applications have been reported in the hydrology and geology literature where selection of sampling or monitoring locations has been approached using probabilistic optimization techniques. Early applications, involving resource assessment generally in mining applications, were based on the assumption that the level of the state variables in the region (ore grades, in many applications) could be modeled as a Gaussian random field using regionalized variable theory. For these applications, the utility of any future sample set could be assessed prior to collection and analysis by estimating the reduction in kriging variance or block kriging variance that the set yielded.

Recently, environmental restoration applications addressing spatial sampling problems have appeared, primarily in the hydrology literature. These applications are often less straight-forward than those based on regionalized variable theory. Most applications are based on flow models that use as input, maps of hydrogeological parameters (porosity or hydraulic conductivity, for examples). Another possible difference between environmental restoration problems and those of the earlier applications is the variety of objectives for different analyses. The objective function determines the utility of further sampling and that of sampling at specific sets of sample locations. The likelihood of a contaminant crossing a site boundary undetected by a specific monitoring system, for example, is an analysis objective that is not related to a reduction in the kriging variance in any obvious way. Consequently, the utility of additional sample or monitoring locations must be approached differently for this objective.

A number of the publications involving spatial sampling problems in environmental restoration focus on the development of methods that facilitate the assessment of state variable information (hydraulic head or contaminant levels, for example) provided by samples of hydrologic parameters. Another primary focus has been in the development of optimization techniques that approach these

specific objectives in an efficient manner. One article that provides an excellent review of accomplishments in these areas and provides numerous further references is Mc Kinney and Loucks (1992). The primary objective in the present paper is to develop a more general approach to optimization for spatial sampling problems that can be applied to a wide range of problems regardless of their modeling requirements and varying analysis objectives.

A sampling plan can be constructed utilizing the appropriate probability space generated using the procedures outlined above, along with optimization methods outlined in this section. A more detailed discussion of these methods is provided in Rutherford (1996). The sampling plan is based on an iterative procedure that requires evaluation of samples collected at each iteration of the process to provide new information for selecting sites to be used in the next iteration. Two issues arise in this process that will be outlined in the next two subsections. The final subsection provides an example of the entire process.

### **Assessing the Value of Sampling at Unsampled Locations**

As alternative sets of supplemental sample locations are considered, there is a need to evaluate the potential for improvement provided by each sample set prior to collection or analyses of the samples. This requires an estimation of the information available from each sample set and an evaluation of its likely impact on the objectives of the overall analysis. Assessment of the relative overall information provided by sampling at any set of sample locations depends on site and modeling information that has been translated to the simulated realizations. The effect of "realizing" values projected by the simulated realizations provides an indication of what might occur if that sample set was actually selected and analyzed.

Figure 2 shows how the resampling process might be implemented. For each hypothetical sample set, realizations (selected at random) are assumed to represent the true status of concentration at the site. Their values at the candidate supplemental sample locations can be submitted as conditioning data along with known sample information to formulate new models. These models can be used to generate a new set of simulated realizations that are then evaluated using the objective function to determine the impact of the new information provided. This process is repeated several times for each supplemental set of sample sites considered. Sets of sample locations that demonstrate only small modeling differences from one realization to the next, or those indicating significant differences to the models, but where the differences have little impact on the objective function, should be excluded from further consideration.

This resampling procedure can be very expensive in terms of computer time, especially when there are flow and transport models that must be reevaluated at each step. Consequently, effective procedures are required for selecting sets of sample sites for consideration.

### **What is Meant by an Efficient Optimization Algorithm**

There are several techniques available for selecting supplemental sample sites. Some techniques will yield an exact optimal solution; other, less computer intensive techniques, will yield a "near optimal" solution. The distinction between these solutions, and the considerations that might provide a preferred technique for any specific application, are outlined in this subsection.

Spatial sampling problems can be approached as a combinatorial optimization problem where a large but finite number of solutions are available. Each possible solution (a set of supplemental sample locations) can be compared through the objective function, to alternatives, using the resampling procedure.

The ideas involved in the sample set selection algorithm are either to find an exact optimal solution dismissing as many solutions as possible without a formal evaluation (recall that formal evaluation may require many executions of a computer intensive simulation involving a flow and transport code), or to find a near optimal solution using the fewest comparisons possible. These methods can be, and probably are, most effectively used in combination.

Simulated annealing and genetic algorithms are methods that compare solutions iteratively, providing a near optimal solution where "near" depends on the efficiency of the algorithm and the number of computational iterations involved. Branch and bound algorithms can be used to find an exact optimal solution. These algorithms derive their benefit (compared to an exhaustive search of every possible solution set) by dismissing large numbers of solutions through a single objective function comparison. Hybrid methods are available that utilize features of all of these random search methods.

Details concerning the concepts and methods involved in these algorithms can be found in references cited in Rutherford (1996). Details of how the procedures might best be implemented for spatial sampling or monitoring problems can also be found in that paper.

## Chemical Waste Landfill Example

This example is provided to illustrate how these spatial sampling procedures might work. The procedure demonstrated here was implemented using a spatial sample tool (OPTMAS), developed at Sandia National Laboratories, for the purpose of locating soil sample locations to provide the maximum relevant site information. The site analyzed is a chemical waste landfill site at Kirtland Air Force Base in Albuquerque, New Mexico. The objective of this analysis (for purposes of illustration) is to determine the extent of a fixed contamination contour with a 90 percent level of certainty.

Figure 3 illustrates the present sample locations by the black dots. Clearly, previous sampling was performed according to some systematic prescription. The center, shape, and extent of the site is presumed to be unknown, but it is assumed that the contamination occurs throughout a contiguous region or, at least, that the contour associated with the outer boundary is the one of interest. This is a realistic problem, for example, in situations where the concern is to achieve a certain level of assurance that a contaminant is contained within site boundaries. It was further assumed that 15 additional sample locations were to be specified at each iteration throughout the analysis. (That is, 15 sample locations are required here. These samples will be taken and analyzed and the process will be repeated if necessary).

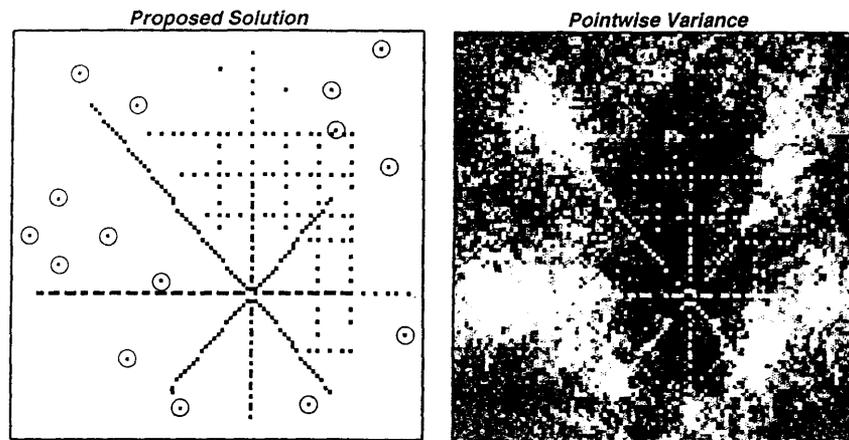


Figure 3

The approach taken to estimate the extent of these contours is the approach described in previous sections of this paper, but with a number of simplifications. First, the simulated realizations are generated using sequential Gaussian simulation (because it was readily available, but not necessarily the best approach to modeling a contaminated region). The set of realizations was used to compute the pointwise variance as illustrated in Figure 3 (the darker areas indicate higher variance). The

reduction in variance is the assumed objective function criterion for this example. Therefore, after the pointwise variance of the set of realizations has been determined, the realizations are no longer needed for the remainder of the analysis (until the next sampling iteration). This further simplification is possible because it is assumed for simplicity that the variance is reduced at, and around, sample locations in a way proportional to the reduction indicated by the previously simulated realizations. The details are provided in Rutherford (1996). This is a luxury that is not offered in the generic spatial sampling problem where new (hypothetical) realizations are required to evaluate each set of potential sample locations. Finally, it was assumed for reasons outside the scope of this paper that the sampling had to occur outside the 10th percentile bound for this contamination contour.

Alternative sample locations were selected using the hybrid sample selection algorithm mentioned above, and detailed in Rutherford (1996). For each sample set selected in the simulated annealing or genetic algorithm portions of the procedure, the objective function (total variance throughout the region) was computed. The sample locations were dismissed or retained as a candidate solution for the problem as prescribed by the hybrid sample site selection algorithm. The circled locations in Figure 3 present a near optimal solution to this problem. It is difficult to determine exactly how close they are to the true optimal solution without using an exact approach. For the present illustration, it suffices to say that the solution appears to satisfy the obvious desirable characteristics of a solution to the problem -- the samples are in areas (constrained as described earlier) where the variability is relatively high and they are spread out throughout this region so that the reduction in variance achieved by sampling at one location will have a minimal effect on the variance reduction achieved by sampling at another selected location.

This simplified example was chosen with a relatively intuitive solution to illustrate the process. The samples selected in a more realistic example might be more difficult to analyze intuitively.

## **References**

Mc Kinney D.C., and D.P. Loucks (1992). "Network Design for Predicting Groundwater Contamination", *Water Resources Research*, Vol. 28, No. 1.

Rutherford, B. (1996). "A General Approach to the Problem of Selecting Optimal Spatial Sampling or Monitoring Locations", white paper (in progress), Sandia National Laboratories, Albuquerque, NM 87106-0829

# Consideration of Spatial Variability in the Management of Non-Point Source Pollution to Groundwater

W. Woldt<sup>1</sup>, F. Goderya<sup>2</sup>, M. Dahab<sup>3</sup>, and I. Bogardi<sup>3</sup>

---

**Abstract.** -- Geostatistical simulation and unsaturated zone modeling are combined to evaluate the impact of spatial variability of selected parameters on groundwater nitrate contamination from agricultural production. Three management scenarios involving spatially variable application were investigated with consideration of spatial variability in residual soil nitrate, yield, and hydraulic conductivity. The process is applied to three conceivable situations, differing in the extent of spatial variability. One decision input, the fertilizer amount, is distinct for different scenarios as well as spatial location for one scenario. Modeling results indicate that variable application of nitrogen, based on spatially variable parameters, not only reduced the over-application of nitrogen, but also reduced the overall groundwater contamination potential. The modeling results indicate that spatial management techniques hold promise for maintaining production while simultaneously protecting the environment.

---

## INTRODUCTION

The natural and management induced variabilities of field soils and hydrologic formations are extensively acknowledged as dominant factors influencing fluid and mass transport through the subsurface zone. The variations of soil and crop properties have led to attempts to understand those variations and to manage production accordingly. The management of production according to localized conditions is variously known as spatially variable, site specific, soil specific, precision, or prescription production. A consensus does not appear to have been reached within the scientific community on the proper term or definition. However, the methodology presented in this research complements production and environmental concerns according to localized conditions and will be referred to as spatial management technology.

---

<sup>1</sup> Assistant Professor, Department of Biological Systems Engineering, University of Nebraska - Lincoln, Lincoln NE.

<sup>2</sup> Graduate Research Assistant, Department of Civil Engineering, University of Nebraska-Lincoln, Lincoln, NE.

<sup>3</sup> Professor, Department of Civil Engineering, University of Nebraska-Lincoln, NE.

The key factor driving spatially-variable control is field variability. Spatial variability may be random or auto-correlated, may be long distance or short distance, and may be small or large. However, all affect the feasibility of spatially-variable control and the design of a particular system to achieve such control. In addition, variability may result from both natural processes and management practices. If there were no variability, production through traditional practice with proper adjustment for field conditions would be adequate. For all its importance, spatial variability and its impact on crop production and the environment has not been studied to a large extent. Furthermore, the procedures for developing an understanding of the variability in a particular field and effective utilization of that information to reduce NPS groundwater pollution, while at the same time maintaining production are still not clearly defined. The purpose of this research is to develop a framework for exploring the effect of spatial variability related to different field scale parameters in managing crop production and minimizing groundwater contamination.

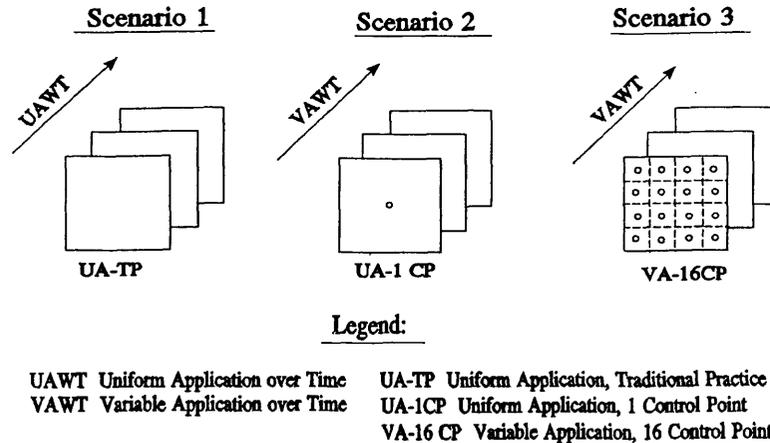
## **METHODOLOGY**

The methodology outlined in this research incorporates spatial variability of various parameters on a field scale in estimating nitrate contamination to groundwater. Unsaturated zone modeling in combination with conventional measurements and simulation is utilized to quantify the effect of field spatial variability on contaminant loading for quantifying environmental benefits of spatial management technology. The approach involves geostatistical simulation to generate a number of realizations reflecting differing degrees of spatial variation. The model is used simulate crop growth and maintain a nitrogen mass balance in the system. Primary elements of the methodology are discussed below.

### **Management Scenarios and Variability**

Three application scenarios were developed in the methodology leading to the spatial management technology. The spatial management scenario is based on appropriate management of nitrogen to meet, but not to exceed production needs. Phosphorus and other nutrients are not assumed to be limiting factors. The scenarios include traditional practices and are modified to reflect possible advancement through application of spatial management technology (Figure 1).

The first scenario is designed to consider uniform input of nitrogen based on traditional practices with the input rate set at a "typical" level and held constant through time. The second scenario assumes the same crop and uniform application, but the amount of application is modified based on soil and yield information from a single location in the field. Thus, the application is uniform over the field, but is variable with respect to time.



**Figure 1. -- Graphical illustration of three application scenarios.**

The third scenario uses variable application based on the spatial variation of selected model and decision parameters. The selected parameters include residual soil nitrate, crop yield and soil hydraulic conductivity. The field is divided into sixteen sectors, and the application rate is based on a single hypothetical measurement of the decision variables in each sector. Hence, the field will receive an application with respect to position as well as time conforming to decision variables within the sector.

The methodology is applied to three different cases which are distinguished by the magnitude of field variability. They are referred to hereafter as; low, medium, and high variability case. In the low variability case, available field information for residual soil nitrate and yield was used to develop the lateral and vertical parameter distributions and geostatistical relations (Goderya et al., 1996; Goderya, 1996). Field conditions were then generated for medium and high variability cases. Mean values for these fields were comparable to the observed field data. However, each case is distinguished by differing coefficients of variation. For this study, values obtained from the literature were used to define these coefficients of variation (Goderya, 1996). Using the subjective prior values of mean, standard deviation and correlation length values for selected parameters, a number of equally likely unconditional realizations were generated. Employing this approach, medium and high variability field conditions were generated for residual soil nitrate and yield values. The general statistical properties for low, medium and high variability fields are given in Table 1.

### **Geostatistical Simulation and Mass Balance**

A total of 100 field realizations, each realization having 120 locations, were simulated for each identified spatially variable parameter under each variability case. Simulated fields were then used as an input to an unsaturated zone transport model for predicting crop production and nitrate loading to groundwater.

**Table 1.--Generated field properties for low, medium, high variability cases.**

	Residual NO3 (kg/ha)			Yield (kg/ha)			Hyd. Cond. (mm/hr)		
	Low	Medium	High	Low	Medium	High	Low	Medium	High
Mean	165.3	146.9	153.9	8965	8580	8378	2.37	2.28	1.29
Std. Dev.	51.5	88.5	163.5	815	2350	3460	0.65	2.08	1.74
Coeff. Var.	0.31	0.60	1.06	0.09	0.27	0.42	0.27	0.91	1.35
Minimum	65.7	29.4	5.2	6300	4156	2112	0.32	0.04	0.01
Maximum	362.5	518.8	861.0	10655	12535	14100	6.69	9.15	9.99

Hence, the complete process of geostatistical simulation and unsaturated zone transport modeling was executed within each variability case for each spatial management scenario. The spatial input data changed between each set of field realizations, but remained constant between each spatial management scenario. On the other hand, decision input changed between each spatial management scenario, but remained identical for various field realizations. For example, spatial inputs of residual soil nitrate, yield, and hydraulic conductivity were different between realizations 1 and 2, but they were identical between realization 1 of scenario 1 and realization 1 of scenario 2 and 3.

A quasi-three-dimensional approach to modeling the important processes and maintaining a mass balance in the system is used in this research. In this case, a one-dimensional model is combined with geostatistical simulation to represent the heterogeneity (spatial variability) in a typical field. The intention of this formulation is to investigate contaminant loading and production from a heterogeneous field within a spatial framework. The model, TDNIT, is used in this research for simulating nitrate movement through root- and vadose- zone (Goderya et al., 1995). However, the methodology is not dependent on the type of model. The use of the selected transport model offers the advantage of short computation time and reduced input data demands. The results from this model also compared well with the program Erosion Productivity Impact Calculator (EPIC) in terms of its predictive ability (Goderya et al., 1995).

A total of 48 simulation runs, each encompassing 100 field realizations, were completed for this effort. The resulting decision input and output were analyzed for each individual model node, and for each run over a five year period.

## CASE STUDY

A case study is presented to demonstrate the methodology. A farm field in Central Nebraska represents a typical crop production area with the potential for nitrate contamination that may be found in the Midwest. Site selection was based upon availability of data, shallow water table, highly permeable vadose zone, rather uniform soil characteristics, and homogeneity of crops produced as well as the general agricultural practices employed. Here, most of the land adjacent to

the site is cropland and nitrogen fertilizer is a major source of crop nutrients and nitrate contamination to groundwater. Continuous corn is selected as the crop to be simulated since it is the major crop grown in the case study area.

Site specific information for the model input variables was compiled from a variety of sources. The transport program required meteorological input data in the form of daily precipitation (including irrigation amounts), average monthly temperature, average monthly solar radiation and albedo. Additional input data included depth to groundwater, potential mineralization amounts, the number of soil layers, soil data for each layer, plant data, and the initial moisture conditions for each layer. The soil input data (porosity, field capacity, permanent wilting point, sand content, organic matter content, saturated hydraulic conductivity, and residual water content) and initial conditions (water content and nitrate content) for the field were based on the vertical profile of the soil horizon and the unsaturated zone. The plant data include crop type, four different nitrate uptake coefficients, leaf area index values and root depth values as a function of time, dry matter/yield ratio, and potential maximum yield. The nitrogen fertilizer data include application amount, date and depth of application.

## RESULTS AND DISCUSSION

The model data were evaluated in terms of area-wide leaching potential and subsequently for leaching as a function of spatial management practice. For all evaluations, output variables were used for relative comparisons. These results represent short-term impacts of the spatial management technology and were calculated for the following variables of interest: annual fertilizer input, annual crop uptake of nitrogen, and nitrogen losses to groundwater.

Figure 2 summarizes the primary nitrogen input. In the low variability case, nitrogen fertilizer input was set at 200 kg/ha/yr (178 bu/ac/yr) for the traditional practice scenario, which compares favorably with the average corn fertilization rate in Nebraska (Follet et al., 1991). Fertilizer nitrogen inputs were lower for second scenario because of the information contained in the field sample which was used to adjust application rate. The overall amount of nitrogen fertilizer applied to the crop in the third scenario was reduced further due to the added level of information and use of spatial application methods.

The simulated crop nitrogen uptake was used as a measure of yield potential. The annual uptake was similar under management scenarios 1 and 2 for all three cases of variability (Figure 3). However, there was a detectable decrease in uptake for scenario 3 (ie., spatial management technology). Based on these results, it appears as though production based on spatial management technology compares reasonably well with the traditional baseline output. Hence, while the spatial management scenario reduced the fertilizer requirements, it also resulted in simulated reduction in the uptake.

The predicted amount of nitrate-nitrogen loading to the groundwater is presented in Figure 4. Potential nitrate loading to groundwater is significantly

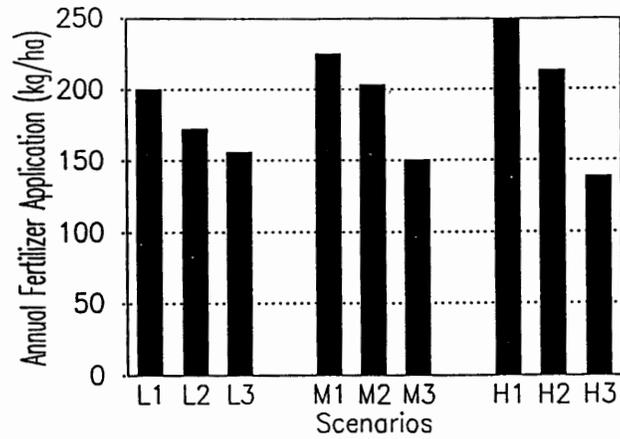


Figure 2. - Annual nitrogen input for variability cases and management scenarios.

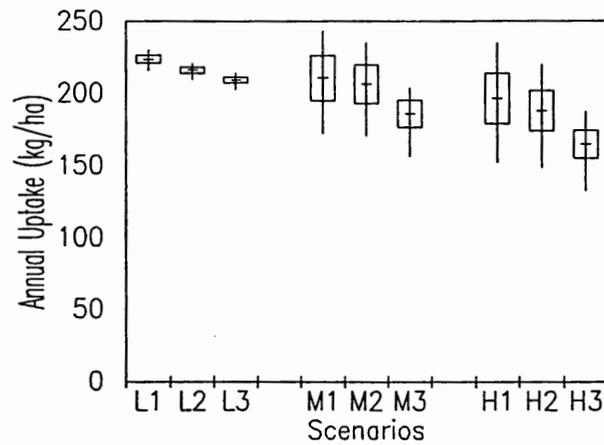


Figure 3. - Simulated annual crop uptake for variability cases and management scenarios.

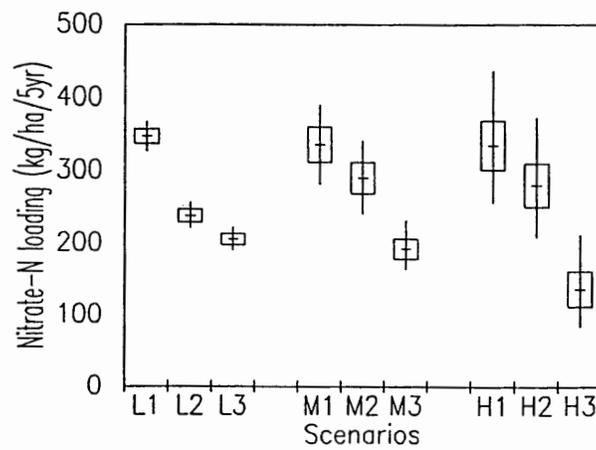


Figure 4. - Simulated nitrate leaching for variability cases and management scenarios.

reduced for all three scenarios. However, the reduction is especially evident in the third scenario with high levels of field variability. The results of these simulations are consistent with the magnitude of the responses reported from experimental studies in the Midwest (Ferguson et al., 1994; Phillips et al., 1993).

These results suggest that increased information for different parameters and conditions in a spatial management system may not necessarily result in a significant decrease in nitrate loading to groundwater. In fact, it appears that the success of spatial management technology is highly dependent on the degree of spatial variability of the primary field parameters that influence nitrate fate and transport. Thus, the effort associated with spatially variable application in a field with inherent low variability of key parameters would most likely be too costly to justify the benefits. However, as the field variability of the key process dependent parameters increases, the apparent environmental benefits of spatial management technology become clearer as depicted in Figure 4.

## **SUMMARY AND CONCLUSIONS**

The methodology presented in this research incorporates spatial variability of various parameters in estimating groundwater contamination potential. It employs a combination of geostatistical simulation and unsaturated zone transport modeling. The methodology outlined in this research was applied to three conceivable variability cases, differing in extent of spatial variability. In all the cases, management practices of spatially variable application were evaluated on the basis of sustaining agricultural production and minimizing environmental pollution. Three management scenarios, including one using spatial management technology were developed and evaluated given the spatial distribution of residual soil nitrates, yield, and hydraulic conductivity in the fields.

The first case was defined as one exhibiting low variability. Results of scenario modeling for this case indicate that while there is a slight reduction in nitrate leaching, the use of spatial management technology may not necessarily result in substantial benefits. It appears that uniform application, given realistic yield goals and composite soil samples, may achieve results comparable to spatial management technology in fields exhibiting low spatial variability. However, use of variable rate application not only reduced the over-application of nitrogen, but also reduced the non-point source pollution to groundwater due to agricultural practices for fields exhibiting both medium and high variability. The methodology described in this study provides a framework to ascertain if the use of spatially variable fertilizer application, or spatial management technology, is environmentally friendly and also provides a basis to evaluate the economics of the technology.

## **ACKNOWLEDGEMENTS**

This paper is supported, in part, by the Center for Infrastructure Research, the Water Center, and the Agricultural Research Division at the University of

Nebraska-Lincoln and, in part, by the Cooperative State Research Service (CSRS) of the U.S. Department of Agriculture.

## REFERENCES

- Ferguson, R.B., G.W. Hergert, J.E. Cahoon, T.A. Peterson, C.A. Gotway, and A.H. Hartford, 1994, "Managing spatial variability with furrow irrigation to increase nitrogen use efficiency," *Second International Conference on Site-Specific Management for Agricultural Systems*, Bloomington / Minneapolis, Minnesota, March.
- Follet, R.F., D.R. Keeney, and R.M. Cruse, 1991, "Managing nitrogen for groundwater quality and farm profitability," *Published by Soil Science Society of America Inc.*, Madison, Washington.
- Goderya, F.S., 1996, "Evaluation and estimation of groundwater potential using spatial parameters," Ph.D. Dissertation, Department of Civil Engineering, University of Nebraska, Lincoln, Nebraska.
- Goderya, F.S., M.F. Dahab, W.E. Woldt, I. Bogardi, 1995, "Comparison of Two Transport Models for Predicting Nitrates in Percolating Water", Submitted for publication in *Transaction of the ASAE*.
- Goderya, F.S., M.F. Dahab, W.E. Woldt, I. Bogardi, 1996, "Spatial Patterns Analysis of Field Measured Residual Soil Nitrate", In *Geostatistics for Environmental and Geotechnical Applications, ASTM STP 1283*, R. Mohan Srivastava, Shahrokh Rouhani, Marc V. Cromer, A. Ivan Johnson, Eds., American Society for Testing and Materials, Philadelphia (In press).
- Phillips, D.L., P.D. Hardin, V.W. Benson, and J.V. Baglio, 1993, "Nonpoint source pollution impacts of alternative agricultural management practices in Illinois: A simulation study", *J. Soil and Water Conservation*, 48:5:449-457.

## BIOGRAPHICAL SKETCH

Wayne E. Woldt is an assistant professor in the Department of Biological Systems Engineering at the University of Nebraska-Lincoln. His interests include risk analysis, spatial management technology, and environmental systems analysis.

Farida Goderya is a graduate research assistant in the Department of Civil Engineering at the University of Nebraska-Lincoln. She is currently completing her Ph.D. dissertation on spatial management technology.

Mohamed F. Dahab is a professor in the Departments of Civil Engineering and Biological Systems Engineering at the University of Nebraska-Lincoln. His interests include pollution prevention and nitrate treatment methods for municipal water supplies.

Istvan Bogardi is a professor in the Department of Civil Engineering at the University of Nebraska-Lincoln. His interests include risk assessment, systems analysis for environmental issues, and global climate change.

# Incorporating Soil Variability into a Spatially Distributed Model of Percolate Accounting

Andrew S. Rogowski<sup>1</sup>

---

**Abstract.**—The Soil-Plant-Atmosphere-Water simulation model (SPAW)<sup>2</sup> was used to compute percolate flux below the root zone from profile water balance. The water balance was based on the prevailing conditions in the atmosphere, physical variables describing water status of soil, and physiological variables related to a stage of growth and water use by plants. Although measured, or estimated at a point, most of the above variables are likely to vary in space and in time. The study addresses this variability and describes how to simulate spatial and temporal distributions of hydrologic variables which are *conditioned* by a set of field observations. We use sequential indicator simulation to convert water flow evaluated at a point in relatively few locations, into spatial flow estimates. The study describes distribution of vertical and horizontal percolate flux in time. It is cast in a GIS format at a catchment scale.

---

## INTRODUCTION

To describe behavior of natural systems we often rely on field observations and numerical models. Observations are generally made at a point, while models attempt to integrate the response spatially and temporally. The field observations considered representative of an area, or volume of soil, serve as input to a model. At other times when field data are not available, modelers use published soil surveys to extract pertinent input variables needed to run a model. Such variables are assumed to apply uniformly to an entire area mapped as the same soil polygon. In truth however, the variables represent observations made at a site, which although typical for a given soil, may be far removed from the study area. Moreover soil polygons themselves are far from pure and contain *inclusions* of associated materials (Rogowski and Wolf, 1994). The objective of this study was to incorporate spatial soil variability expressed as different descriptions of the same soil into a simulation model of percolate accounting.

---

<sup>1</sup> Soil Physicist, USDA-ARS-PSWMRL, University Park, PA.

<sup>2</sup> The SPAW model was developed by Dr. Keith Saxton USDA-ARS Pullman, WA.

## METHODS AND MATERIALS

### Soil-Plant-Atmosphere-Water (SPAW) Model

The SPAW simulation model (Saxton et al., 1992) provides estimates of evapotranspiration (ET), plant water stress and layer-by-layer changes in soil water content and percolate flux from below the root zone. The model is designed to provide a daily, one dimensional (vertical) water budget at a point. Actual measurements of ET, soil water content and other input variables may be substituted for default values, or periodically used to validate, or recalibrate the model. The model requires climatic and plant data and a detailed *in situ* horizon by horizon description of the soil profile, including pertinent soil water characteristics, or textural classification information for each soil layer.

Potential evapotranspiration (PET) is determined from measured values of pan evaporation. The PET values are used to compute actual ET which is then subtracted from the available soil water in storage. Daily precipitation, corrected for interception and runoff, infiltrates and is distributed among soil layers by *cascading* down from the surface. The amount of water which reaches the bottom boundary layer becomes a deep vertical percolation flux ( $q_v$ ). Adjustments among layers due to unsaturated water flow by redistribution are made using a Darcy subroutine.

The output consists of daily summaries of soil water for each layer, in addition to a digital and graphical output of precipitation, ET, and percolation flux  $q_v$ . The major feature of the model is that it addresses the soil-plant-air continuum at a point, and estimates are made conditional to the profile hydrologic properties, subject to the local climatic and biotic factors.

### Study Area and Soils

The study area described in this application is on the Mahantango Creek watershed in Pennsylvania, USA, (Rogowski and Wolf, 1994). The watershed is situated in the Valley and Ridge physiographic province of the Appalachian Mountains, characterized by varying relief: upland hills, valleys, and forested mountain ridges dissected by streams. The climate is humid with well-distributed rainfall (about 1000 mm/yr.). Elevation ranges from less than 100 m in the valleys to over 500 m on the ridge tops. The ridges, valleys and streams are oriented northeast-southwest, corresponding to the regional strike of major rock formations, while beds dip to the northwest and southeast along the centrally located anticline. The land use is predominately cropland in a rotation management system of corn, small grains, and meadow, intermingled with numerous tracts of woodland, areas of permanent pasture and orchards.

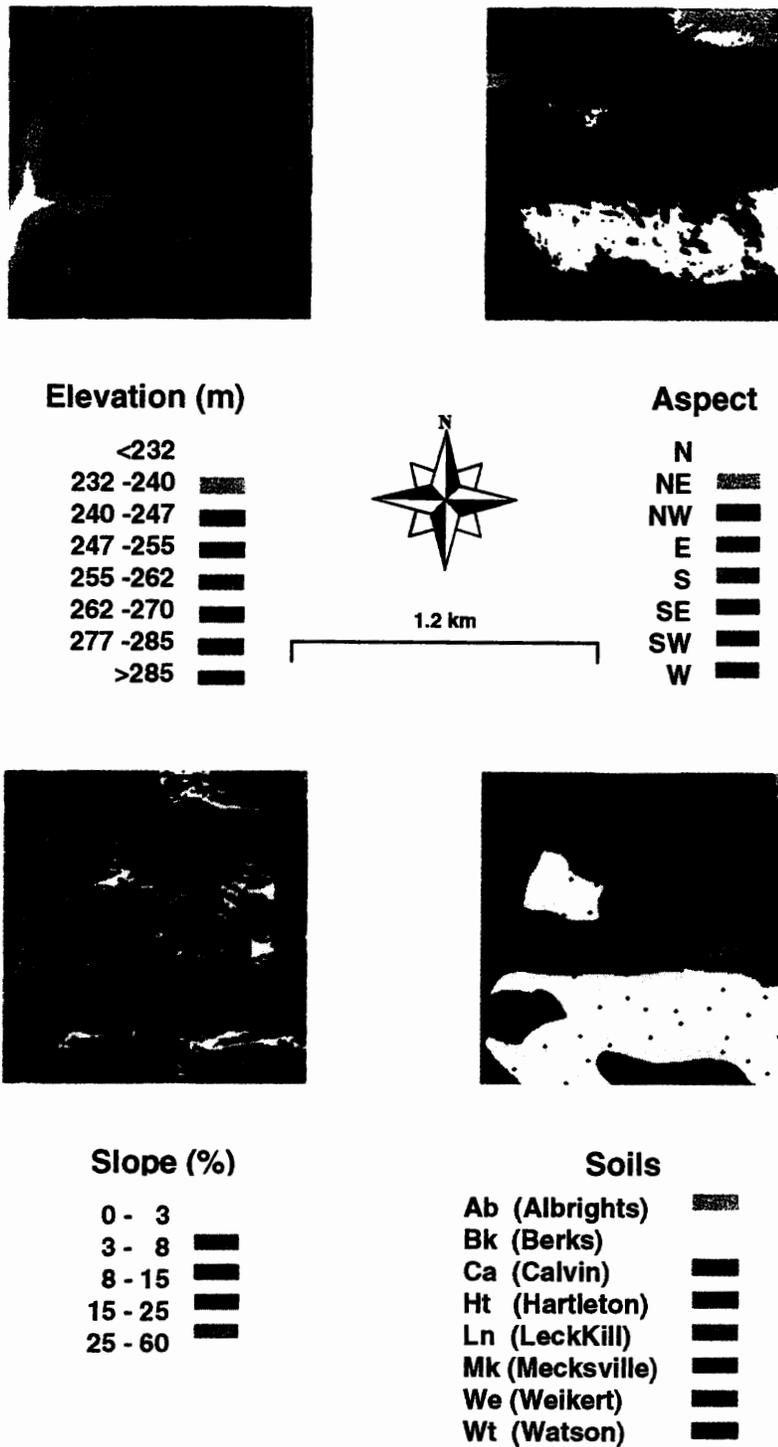


Figure 1. Overlay maps of elevations, aspect, slope, and soils based on the detailed 5m DEM of the study area; locations of conditioning pedons are shown on the soils' map.

The experimental area is 1.2 km on the side with the individual cell size of 5×5 m corresponding to a detailed DEM. Figure 1 shows distributions of georeferenced (GPS) elevations, slopes, aspects, and soils at the experimental site. Points on the soils map indicate *assigned* locations of the *soil pedons* used as input to the SPAW model. The displays and calculations were carried out within the framework of GIS IDRISI<sup>3</sup> (Eastman, 1993).

### Soil Profiles

Soil profile locations (*pedons*) were assigned at random, using as input the information from PSU Soils Database (Ciolkosz and Thurman, 1992). This database is a compilation of 800 pedons representing 170 different soil series which have been collected and analyzed by the soil characterization laboratory. It contains detailed profile and site descriptions, as well as layer-by-layer physical, chemical and mineralogical data. There are generally several descriptions of the same series sampled at different locations. For the eight soils found in the experimental area 60 pedons were described in the database, these were augmented with pedons of a different series considered as inclusions, for a total of 100 pedons. The pedons, assigned at random to soil polygons of the same series (figure1), constituted the *conditioning data sites*.

### Conditioning Data

At each of the conditioning sites we extracted from the soils database the textural information, as well as percent sand, silt, and clay, and depth to each layer. We used daily precipitation and PET data from the 1990 season, and assumed a uniform crop cover (corn) over the entire area. The SPAW model was run on each profile from Spring to Fall. The model was started with a uniform water content ( $\theta_v$ ) of 0.20 cm<sup>3</sup>/cm<sup>3</sup> and recalibrated to volumetric field capacity for each layer.

### Soil Anisotropy

Field saturated hydraulic conductivity  $K_{fs}$  of each layer was estimated from the ten textural classes in the SPAW supplied tables. The  $K_v$  (vertical) for the soil was computed as a harmonic mean of component layers weighted by a layer thickness. The  $K_h$  (horizontal) for the soil was taken as the conductivity of top layer, if followed by a layer with lower  $K$ . If  $K$  in the following layer(s) was higher,  $K_h$  was taken as an arithmetic average of component layers weighted by the layer thickness. The ratio,  $n = K_h / K_v$  was assumed to represent profile anisotropy (Zaslavsky and Rogowski, 1969). Given percolation flux  $q_v$  computed from the SPAW, horizontal flow component  $q_h$  was approximated as,

$$q_h = q_v \times n \times \tan \alpha \quad [1]$$

where  $\alpha$  represents land slope in % (figure 1) derived from a detailed 5 m DEM, and  $\tan \alpha$  is equivalent to (slope gradient)<sup>-1</sup> definition in IDRISI<sup>3</sup>

<sup>3</sup>The mention of trade names does not constitute an endorsement of the product by the US Department of Agriculture.

## Sequential Indicator Simulation

Results in literature (Journel and Alabert, 1989), suggest that a geostatistical conditional simulation approach known as sequential indicator simulation (SISIM) can be used to reproduce spatial patterns of continuity and to introduce a measure of uncertainty into soil survey data. The purpose of the simulation is to construct from the conditioning data (100 profiles), numerous, equiprobable spatial “realizations”, or outcomes, of an attribute  $q_v$ , that would reproduce spatial patterns associated with the occurrence of extreme values.

To model different portions of an irregular distribution, the sequential indicator simulation partitions observed population into several classes each separated by a threshold. For each observed value it then defines the indicator transformation with respect to that threshold. Thus for a random variable  $Z(\mathbf{u})$  at a location  $\mathbf{u}$  and a threshold value  $z$ , the corresponding indicator transformation  $I(\mathbf{u};z)$  would be,

$$I(\mathbf{u};z) = \begin{cases} 1, & \text{if } Z(\mathbf{u}) \leq z \\ 0, & \text{otherwise} \end{cases} \quad [2]$$

Given a set of  $q_v$  values, the SISIM procedure converts them into indicator distributions for each threshold. It then attempts to define an algorithm for adding an indicator value at an unsampled location which is consistent with a spatial structure (variogram or covariance) of observed population for a particular threshold (Deutsch and Journel, 1992). The first step involves estimation of the conditional probability that a new value is less than a given threshold. This is done by kriging each unsampled location using the surrounding indicator values. The resulting estimate of the conditional cumulative distribution function (ccdf) is between 0 and 1. The actual simulation of a corresponding indicator value at an unsampled grid location is accomplished by drawing a random number between 0 and 1 from a uniform distribution (Monte Carlo). If this number is less than, or equal to a kriged value, a simulated indicator value of 1 is assigned to that location. If it is more, a value of 0 is assigned. The new value becomes a part of the conditioning data set and the procedure is repeated at another location until all unsampled locations in a catchment grid are filled.

## RESULTS

### Distributed Flow

After generating 25 equiprobable realizations of cumulative percolate flux  $q_v$ , expected values and quantiles of the average of 25 distributions were analyzed for three dates: March 30<sup>th</sup>, August 15<sup>th</sup>, and November 15<sup>th</sup>, 1990. We show these in figure 2. We also generated 25 realizations of  $n = K_h/K_v$ . Subsequently, expected values of  $n$  and  $q_v$  at each pixel, and slope from the overlay of slope in figure 1 were used in Eq. 3 to compute distributions of the horizontal flux component  $q_h$

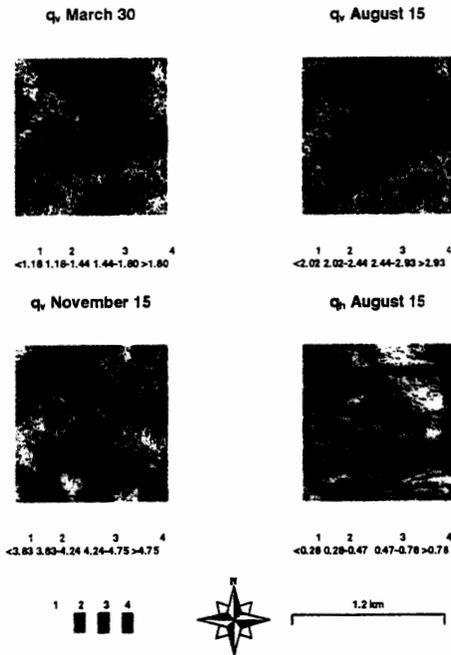


Figure 2. Simulated (SISIM) quantiles of expected values of  $q_v$  (cm) on Mar 30<sup>th</sup>, Aug 15<sup>th</sup>, and Nov 15<sup>th</sup> and of  $q_h$  (cm) on Aug 15<sup>th</sup> based on 100 conditioning data.

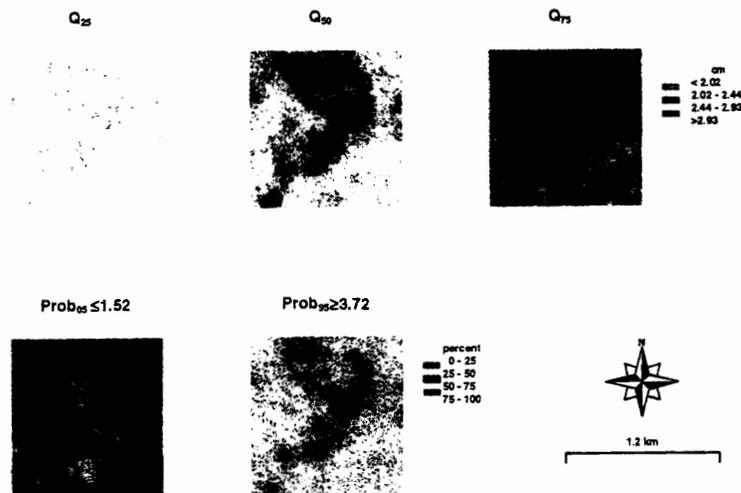


Figure 3 Distribution of  $q_v$  (cm) for three quantiles (Q) of 25 SISIM simulations on Aug 15<sup>th</sup>; and probability (Prob) of values lower than  $Q_5$  (1.52 cm), or higher than  $Q_{95}$  (3.72 cm).

for each location and date. Since there was not much difference among the three dates, figure 2 shows  $q_h$  for August 15<sup>th</sup> only. We represent the  $q_v$  and  $q_h$  as quartiles (Q) of the *average* (25 SIS) distribution: (1) *yellow*  $< Q_1$ , (2) *green*  $Q_1$  to  $Q_2$ , (3) *blue*  $Q_2$  to  $Q_3$ , and (4) *red*  $> Q_3$ . Results suggest that areas with high  $q_v$  (red) may be the sites of preferential recharge. Similarly locations with above average  $q_h$  may indicate preferential interflow, or runoff producing zones. The values of  $q_h$  are very dependent on interaction among the distributions of slope (from 5 m DEM), and distributions of  $n$  and  $q_v$  (from two separate 25 SISIM runs).

The overall distribution of  $q_v$  does not differ much from season to season. In the Spring a very prominent *red* zone occurs the middle of the catchment. As time goes on, this zone becomes more diffuse in the Summer, and brakes into several patches in the Fall. When results for  $q_v$  in figure 2 are compared with corresponding terrain properties in figure 1, the primary recharge area seems to be in the center of the catchment. This is also zone of terrain depression and a natural flow-pathway which connects part of the catchment to the main drainage. Ironically, the farm buildings are located in this same general area.

### Quantile Maps

To derive a quantile map, 25 simulated values for each pixel were ranked and the quantiles (Q), were selected as simulated pixel values corresponding to the 25<sup>th</sup>, 50<sup>th</sup>, and 75<sup>th</sup> percentile in the ranking. These pixels were subsequently combined into a single quantile overlay. The three quantiles (figure 3) were  $Q_{25}$ , median ( $Q_{50}$ ), and  $Q_{75}$ . The  $Q_{25}$  realization focuses on highlighting the potential outliers, since the probability of a lower value than shown is only 25%. The median ( $Q_{50}$ ) quantile map tends to minimize the mean of absolute deviations for each pixel location, that is, individual pixel values are close to what they appear to be. In the  $Q_{75}$  map the focus shifts to the reliable low zones, because in this case the probability of exceedance would be only 25%. This map and the  $Q_{50}$  map are the closest in appearance to that in figure 2. The results suggest that the *red* zone is real and its occurrence persists in the quantile maps. Equally persistent are the low flow (*yellow*) areas in the east and south east corner and along the lower portion of western boundary.

### Probability Maps

Probability maps (figure 3) provide additional detail about the existence and continuity patterns associated with either the low, or high zones of  $q_v$ . The maps are expressed as percent probability that an area is either higher, or lower than a given cutoff. The choice of a cutoff is arbitrary. Our maps were constructed for two cutoffs:  $\text{Prob} \leq Q_5$  and  $\text{Prob} \geq Q_{95}$ . The  $P \leq Q_5$  map delineates potential low flow zones, while the  $P \geq Q_{95}$  map shows high flow areas in terms of the probability of occurrence. Both maps confirm the existence of centrally located, connected, high  $q_v$  area, and corresponding low flow zones in the eastern and western portions of the catchment.

## CONCLUSIONS

To introduce a measure of spatial variability into soil survey data, we extracted detailed pedon information from a soils' data base and assigned it at random to locations within a mapped polygon representing that soil. The assigned pedon locations became the sites of conditioning data. We run a model (SPAW) to compute daily water balance for each pedon, extracting the percolate flux component  $q_v$ . The values of  $q_v$  were then used as conditioning data in the sequential indicator simulation model to distribute  $q_v$  over the catchment. Results suggested the existence of high and low flow zones reflecting terrain properties. Their continuity was confirmed by the quantile and probability of exceedance maps. An added benefit was the identification of horizontal flux component  $q_h$ , related to soil anisotropy. Distributions of  $q_h$  were strongly slope dependent and reflected contributions by the percolate flux  $q_v$  and local anisotropy  $n$ .

## ACKNOWLEDGMENTS

Contributions by D. Simmons to data visualization and processing are appreciated.

## REFERENCES

- Ciolkosz, E.J. and Thurman, N.C. 1992. Soil Char. Lab. Database System. Agron. Ser. No.124. Agron. Dept., Pen State Univ., University Park, PA 16802. 59p.
- Deutsch, C.V. and Journel, A.G. 1992. GSLIB, Geostatistical software library and users guide. Oxford Univ.Press, New York, NY.
- Eastman, J.R. 1992. IDRISI a grid based geographic analysis system. Version 4.1. Clark University Graduate School of Geography, Worcester, MA 01610. 178p.
- Journel, A.G and Alabert, F. 1989. Non-Gaussian data expansion in earth sciences. Terra Nova 1: 123-134.
- Rogowski, A.S. and Wolf, J.K. 1994. Incorporating variability into soil map unit delineations. Soil Sci.Soc.Am.J., 58:163-174.
- Saxton, K.E., Porter, M.A., and McMahan 1992. Climatic impacts on dryland winter wheat by daily soil water and crop stress simulations. Agric. and Forest Meteorology, 58:177-192.
- Zaslavsky, D. and Rogowski, A.S. 1969. Hydrologic and morphologic implications of anisotropy and infiltration in soil profile development. Soil Sci.Soc.Am.J., 33(4): 594-599.

## BIOGRAPHICAL SKETCH

Andy Rogowski is a Soil Scientist at the USDA-ARS Pasture Systems and Watershed Management Research Laboratory, he graduated from the Iowa State University at Ames with a Ph.D. degree in Soil Physics.

# Response of a Watershed Model to Varying Spatial Landscape Characteristics

Dennis P. Swaney<sup>1</sup>, Wen Ling Kuo<sup>2</sup>, David A. Weinstein<sup>3</sup>, Charles Mohler<sup>4</sup>, Stephen DeGloria<sup>5</sup>, Chris Pelkie<sup>6</sup>, Fuan Tsai<sup>7</sup>, Tammo S. Steenhuis<sup>8</sup>, and Charles E. McCulloch<sup>9</sup>

---

**Abstract.** Using a spatially-explicit deterministic model of hydrology and soil nitrogen dynamics in a small watershed in upstate New York, we investigate the effect of varying the spatial distribution of land use on processes within the watershed. Multiple landscapes having specific properties are generated, and the effects on runoff of water, runoff loss of nitrogen, and other soil nitrogen processes are observed over a period of two years. We illustrate techniques of visualization for comparing total magnitude and spatial and temporal distributions of these variables. The ultimate goal of our approach is to develop a methodology for use in comparative studies of land use modification at the landscape scale.

---

## INTRODUCTION

Because nitrogen can be seen both as a pollutant and a nutrient, understanding the nitrogen dynamics of watersheds is crucial to both surface and groundwater quality management, to effective agricultural management, and to the advancement of our understanding of other terrestrial and aquatic ecological processes. With the explosive growth of computer resources over the last few decades, mathematical modeling of nitrogen at the watershed scale has advanced from empirical loading models, to process-based lumped-parameter models, to spatially explicit models. While the first two categories of model permit the estimation of annual or seasonal loading rates, and the construction of average mass balances of some watersheds, they fail for complex watersheds in which flows are generated primarily in small "source areas". They also do not allow examination of the impacts of the position of

<sup>1</sup> Extension Associate, Center for the Environment, Cornell University, Ithaca, NY

<sup>2</sup> Graduate Student, Agricultural and Biological Engineering, Cornell University

<sup>3</sup> Forest Ecologist, Boyce Thompson Institute for Plant Research, Ithaca, NY

<sup>4</sup> Senior Research Associate, Ecology and Systematics, Cornell University

<sup>5</sup> Associate Professor, Soil, Crop & Atmospheric Sciences, Cornell University

<sup>6</sup> Scientific Visualization Producer, Center for Theory and Simulation in Science and Engineering, Cornell University

<sup>7</sup> Graduate Student, Civil and Environmental Engineering, Cornell University

<sup>8</sup> Associate Professor, Agricultural and Biological Engineering, Cornell University

<sup>9</sup> Professor, Biometrics Unit and Statistics Center, Cornell University

nitrogen sources in the watershed, and the effects of correlations between sources and topography and soil properties which affect nitrogen transport and transformations. Even some of the most successful watershed models which incorporate spatial information do so by aggregating topographic or soil data (e.g., Beven and Kirkby, 1979) or by using monthly or greater time steps (e.g., Bartell and Brenkert, 1990), thereby losing the detail of the spatial interactions and the response to individual rainfall events.

Under what circumstances should these details matter? If we are considering alternative strategies for managing a watershed in which we must decide the location of nitrogen sources (e.g. cornfields) it may be of interest to know whether location strongly affects nitrogen load from the watershed. A related question is whether an existing mosaic of land use in a watershed represents the smallest environmental impact in terms of surface or groundwater loading, or if alternative mosaics could be more desirable.

Below, we report some preliminary results of our analysis of a small upstate New York watershed which incorporates data describing the spatial distribution of topography, land use, and soils. We discuss the effects of changing land use characteristics on soil water and nitrogen spatial distributions and losses from the watershed in runoff.

## METHODS

### Model Descriptions

#### Hydrology

For this study, we simulate the hydrology and nitrogen dynamics of a 15.3 ha watershed in upstate New York on a daily time step for two years, beginning in April, 1991. We use a GIS-based hydrology model originally developed by Zollweg and Steenhuis (Zollweg, 1994) and adapted for our application (Kuo et al., 1996) to drive a spatially explicit soil nitrogen model. Hydrology is driven by local daily precipitation and pan evaporation data

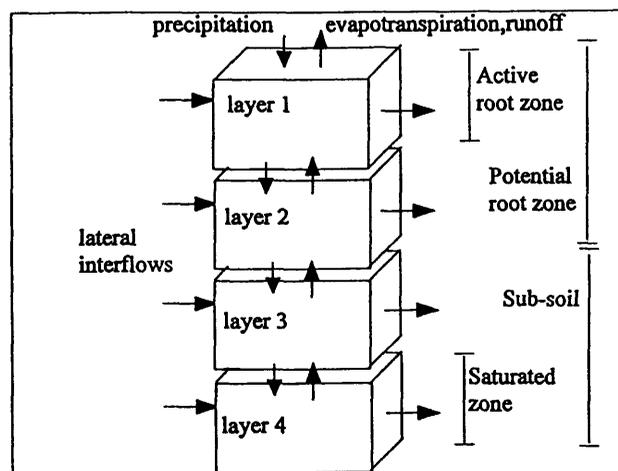


Figure 1. Schematic of hydrology model

obtained from the Northeast Regional Climate Center at Cornell. The hydrology model describes the watershed as a collection of 1531 100 m<sup>2</sup> cells which are divided into 4 layers (figure 1).

The top layer represents the active root zone the bottom boundary of which varies in depth through the growing season. Layer 2 is the zone between this depth and the maximum depth of the root zone. Layer 3 represents the unsaturated zone below the potential root zone, and is bounded below by the water table. Layer 4, the saturated zone, is bounded below by impervious bedrock, the depth to which depends on local soil type.

Water enters the top layer of each cell through rainfall and snowmelt. Lateral exchange of water between a cell and its eight neighbors is determined by the product of local hydraulic conductivity and elevational gradient. Vertical exchanges are governed by mass balance between layers, given the impervious bottom boundary condition of the bottom layer. Water in excess of saturation is lost from the top layer as surface runoff. Water is also lost from the active root zone by evapotranspiration.

### Nitrogen

The nitrogen model (figure 2) includes exchanges between three components in each cell: organic N, ammonia, and nitrate. Nitrogen is added to the watershed in the form of dry deposition (daily) and wet deposition (episodically with rain) of nitrate and ammonium, and through fertilizer applications of nitrate on agricultural lands. In the simulations discussed here, a 100 kg/ha application of nitrate to crop land occurs on May 31 of each year. Exchanges between the N pools are mediated by microbial activity which depends on temperature and soil moisture content.

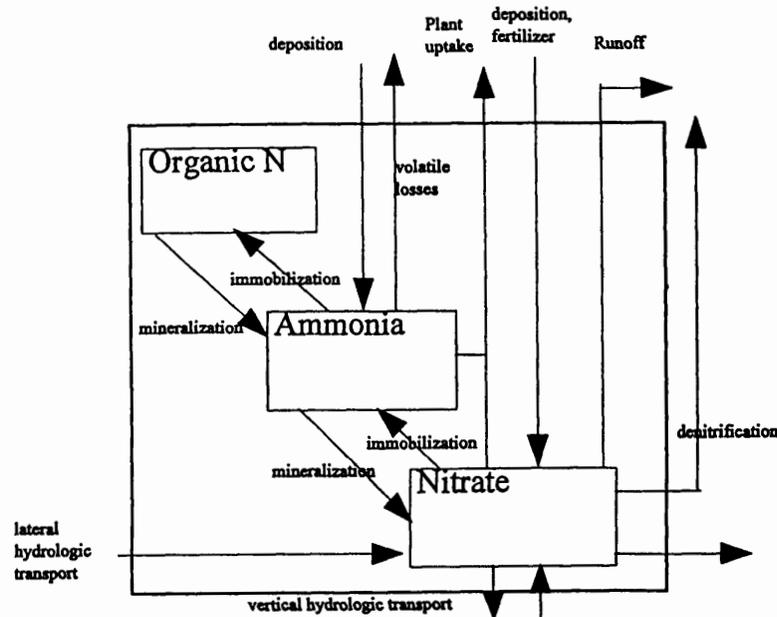


Figure 2 . Schematic of nitrogen model

Organic nitrogen may be mineralized to ammonia, which in turn may be nitrified or lost to the atmosphere through volatilization. Nitrate is lost to the atmosphere by denitrification under wet soil conditions. Nitrate is assumed to be the only mobile component of nitrogen. It moves in dissolved form with water within the watershed, and is lost from the watershed in runoff water. Both ammonium and nitrate are taken up by plants following a demand schedule determined by the land use category of the particular cell.

### **Landscape**

Land use data characterizing the watershed were determined from 1991 aerial photographs (1:24,000; figure 3a); digital elevation data were digitized from contours of USGS 1:24,000 topographic maps (figure 3c); soil maps (1:20,000) from the Tompkins County soil survey (figure 3d) (Poiani et al. , submitted).

We consider three alternative land use mosaics for the watershed: existing land use ("original"), 100% active cropland ("agricultural") and a "synthetic landscape" generated using a conditionally independent hierarchical model (Searle, et al., 1992) ("random"). The existing land use mosaic includes five categories: active agriculture (43%), forest (41%), wetland (9%), inactive agriculture (4%), and low density residential (3%) (figure 3a). The agricultural mosaic was included to examine the limiting case in which a nitrogen fertilizer is applied to the entire landscape. Detailed description of the synthetic landscape generator is beyond the scope of this paper, but is provided in McCulloch et al. (1995). Briefly, the land use categories are modelled as a function of slope, and are generated using a spatially autoregressive correlation structure to induce a high correlation between land uses in adjacent cells. Synthetic landscapes are useful for assessing variation and for "filling in" landscapes which may undergo change. The example realization of the synthetic landscape model considered here resulted in the following distribution of land uses: active agriculture (54%), forest (22%), wetland (6%), inactive agriculture (1%), and low density residential (18%) (figure 3b).

## **RESULTS AND DISCUSSION**

Hydrology and nitrogen were simulated for two years beginning April 1. Here, we compare the water and nitrate conditions in the landscape at the time of the fertilizer application to cropland in the second year, after any effects associated with the initial condition (saturated soil water content; constant N concentrations) have dissipated, and again just prior to the start of the next growing season.

Maps of the total soil moisture content for the "original" mosaic are shown in figure 3e and 3f. We show the results from this case only because those corresponding to the other mosaics are nearly indistinguishable from this one. The figures indicate that by the end of May in the second year of the simulation, and persisting through and after the growing season, the watershed

has established a distinctive pattern of high moisture content along a central drainage which curves to the lower left where the outlet is located. Even during the driest periods of the simulation, this central feature exists, expanding and contracting following rainfall and snowmelt events.

The average nitrate concentration in the root zone corresponding to the same dates for each of the land use mosaics are shown in figures 4a and 4b, along with cumulative runoff losses of nitrate. For all three landscape mosaics, the pattern of nitrate concentration follows the distribution of crop land immediately following fertilizer application, as expected, because the rate of nitrate applied is at least an order of magnitude greater than that from typical atmospheric deposition. However, about nine months after application, the pattern shows the effect of water content in the root zone. In particular, the wet areas that occupy the center of the map and curve to the lower left appear to be largely depleted of nitrate. This may be due to a combination of denitrification and runoff losses, both of which occur at their highest rates under saturated conditions. We believe the transitional pattern from a land use dominated mode immediately following a large application of nitrogen to a soil/topography based pattern as hydrological processes proceed is a general feature of mixed use watersheds.

Not surprisingly, cumulative runoff losses exhibit a sharp increase immediately following fertilizer application on all landscapes, with a gradually increasing trend thereafter. The highest losses occur from the "agricultural" mosaic. The "random" mosaic generates nitrate losses intermediate between the "agricultural" and the "original" mosaic.

Of some interest is the relatively high loss of nitrogen in runoff from the synthetic landscape. High losses are also observed in two alternative realizations of the synthetic landscape not shown here. Expressed per area of crop land, the losses from the synthetic landscapes are near or above those generated from the "agricultural" mosaic. One interpretation of this result is that a more sophisticated landscape generation algorithm should take persistent features of soil moisture (or its governing variables: topography, porosity and hydraulic conductivity) into account. It seems unlikely that the factors which have selected the location of cropland in a landscape reduce the runoff losses of nitrogen, but further work with synthetic landscapes may shed light on this issue.

#### **ACKNOWLEDGEMENTS**

Visualizations of the simulation results reported here were obtained using the resources of the Cornell Center for Theory and Simulation in Science and Engineering. The Cornell Theory Center receives major funding from the National Science Foundation and New York State, with additional support from the Advanced Research Projects Agency, the National Center for Research Resources at the National Institutes of Health, IBM Corporation and members of the Corporate Partnership Program.

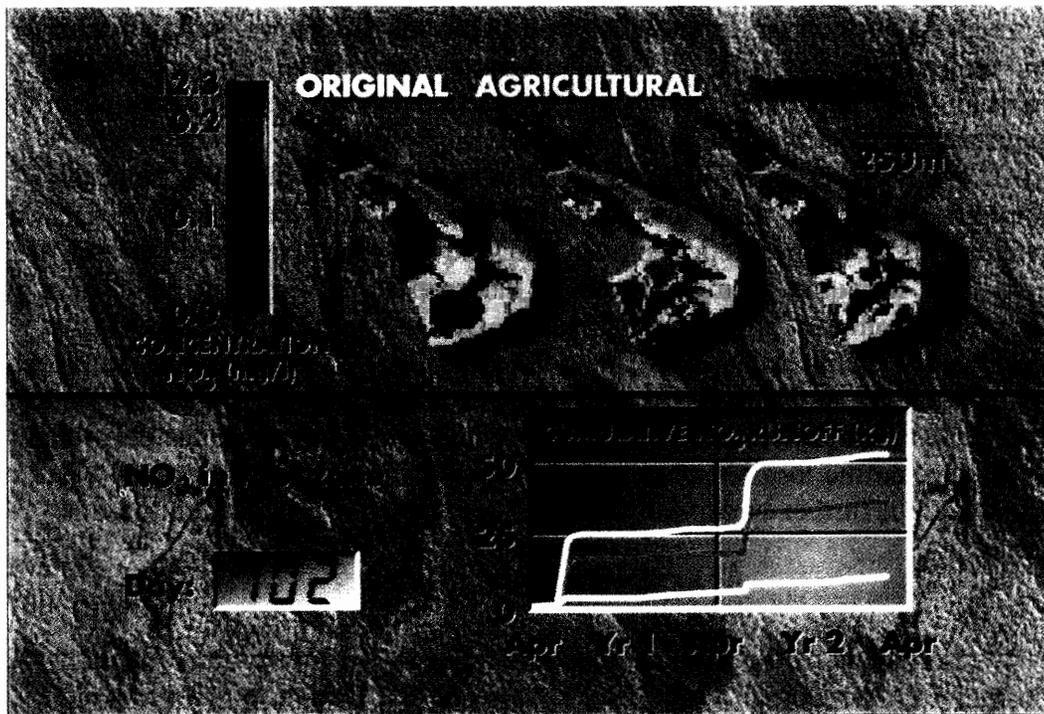
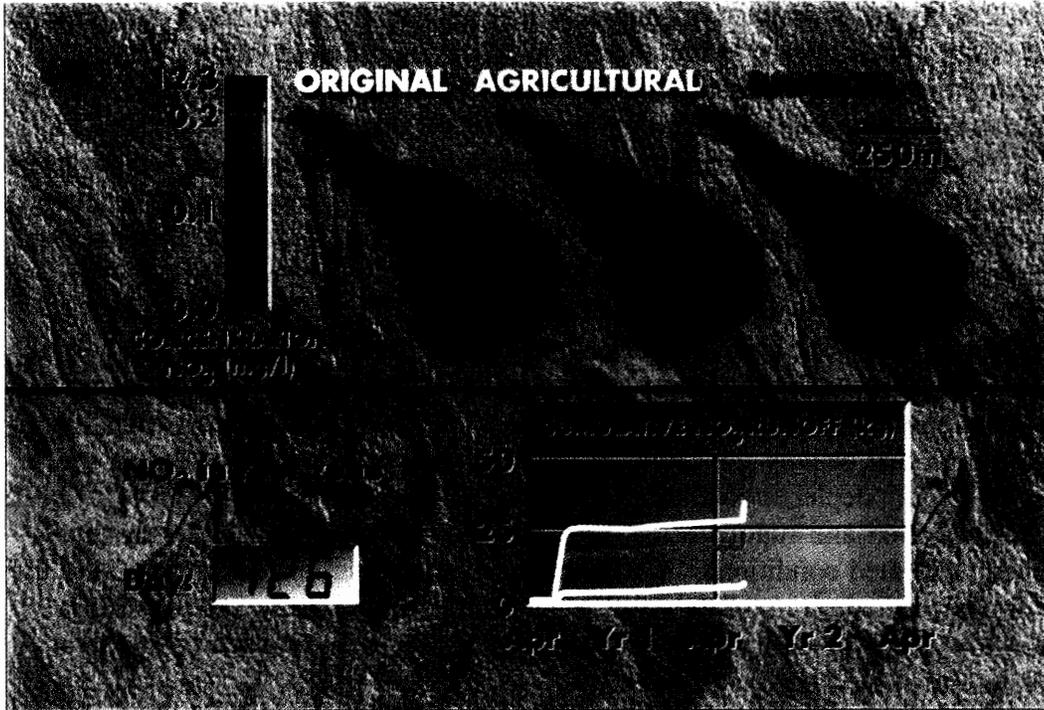


Figure 4 (a) Distribution of nitrate nitrogen in the root zone on the date of the second fertilizer application (day 426) and corresponding cumulative nitrate losses in runoff for the three land use mosaics; (b) same variables 276 days later.

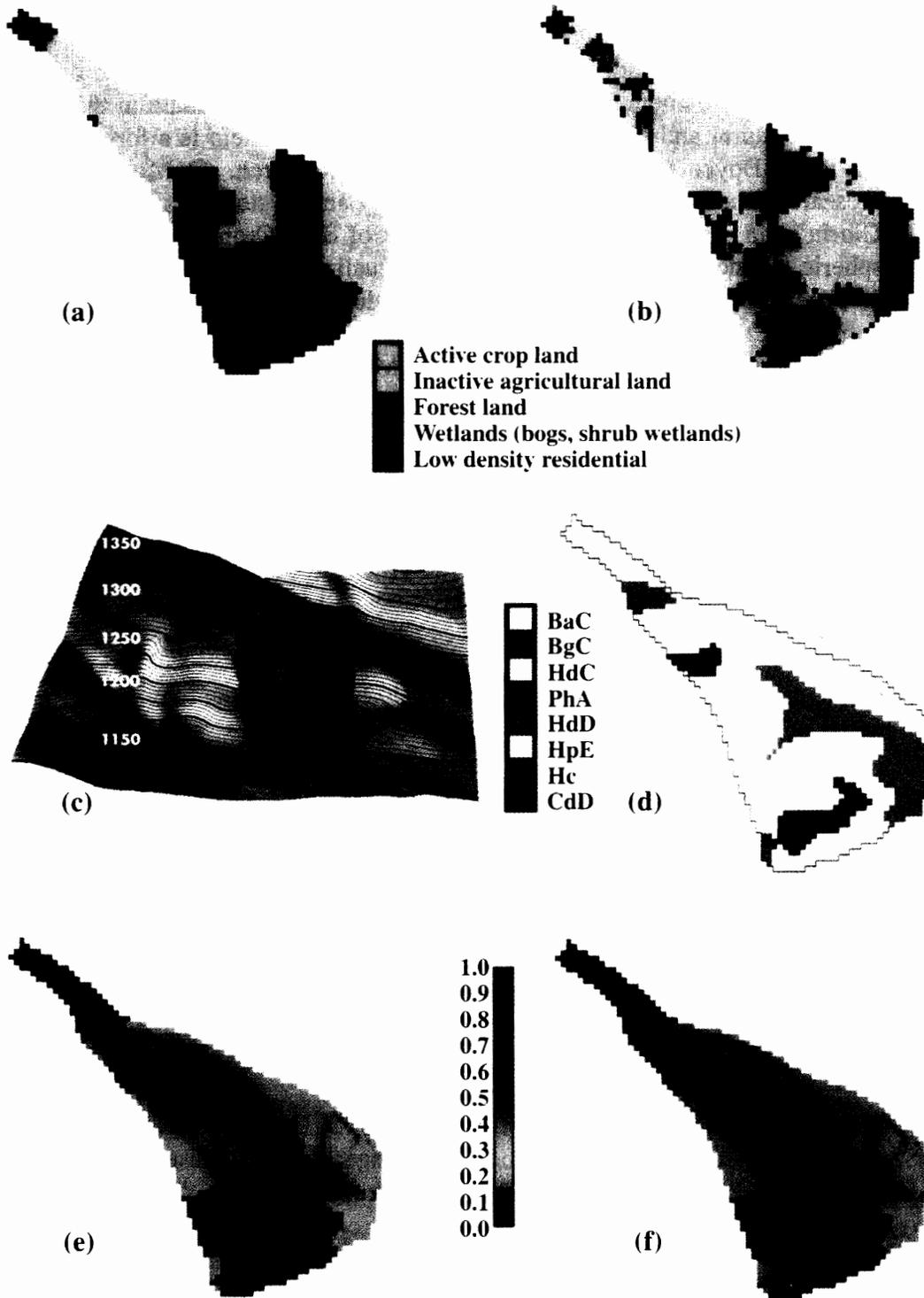


Figure 3 (a) Original land use map, (b) "random" land use map, (c) elevation map with original land use map superimposed and 5' elevation contour lines, (d) soil map, (e) moisture saturation index at time of second fertilizer application, (f) saturation index 276 days later

## BIOGRAPHICAL SKETCHES

D. Swaney is a mathematical modeler in the Biogeochemistry Program of the Center for the Environment at Cornell. W. Kuo is a Ph.d. candidate in the Dept of Agricultural and Biological Engineering. D. Weinstein is a forest ecologist at the Boyce Thompson Institute for Plant Research. C. Mohler is a senior research associate in the Section of Ecology and Systematics at Cornell. S. DeGloria is an associate professor in the Dept of Soil, Crop and Atmospheric Sciences. C. Pelkie is scientific visualization producer at the Cornell Theory Center. F. Tsai is a graduate student in the Dept of Civil and Environmental Engineering. T. Steenhuis is an associate professor in the Dept of Agricultural and Biological Engineering. C. McCulloch is a professor in the Biometrics Unit and Statistics Center.

## REFERENCES

- Bartell, S.M. and A.L. Brenkert. 1991. A Spatial-temporal model of nitrogen dynamics in a deciduous forest watershed., pp 379-398 in: Turner, M. G. and R.H. Gardner (eds.) 1991. Quantitative Methods in Landscape Ecology Ecological Studies 82, Springer-Verlag, NY.
- Beven, K.J. and M.J. Kirkby. 1979. A physically-based variable contributing area model of basin hydrology. *Hydrol. Sci. J.* , 24(1):43-69
- Johnsson, H. 1990. Nitrogen and water dynamics in arable soil. Swedish University of Agricultural Sciences, Department of Soil Sciences Reports and Dissertations, 6. Uppsala, Sweden.
- Kuo, W-L., P. Longabucco, M.R. Rafferty, J. Boll, and T.S. Steenhuis. 1996. An integrated GIS-based model for soil water and nitrogen dynamics in a New York City watershed. Proc. AWRA Symposium "Watershed Restoration Management" (July 14-17, 1996). Syracuse, New York (Accepted)
- McCulloch, C.E., W.L. Kuo, and D.P. Swaney. 1995. Designer landscapes: Simulating landscapes for management of change. Invited Presentation, Eastern North American Region Biometrics Meeting. Birmingham, AL
- Poiani, K. A., B.L. Bedford and M.D. Merrill. 1996. A GIS-based index for relating landscape characteristics to potential nitrogen leaching in wetlands. *Landscape Ecology* (submitted).
- Zollweg, J.A. 1994. Effective use of geographic information systems for rainfall-runoff modeling. Unpublished dissertation, Cornell University.

# Understanding the Spatial Distribution of Tree Species in Pennsylvania

Rachel Riemann Hershey<sup>1</sup>

---

**Abstract.**--Current, accessible information on the distribution of tree species would aid in the understanding and management of ecosystems. However, such detailed information on forest composition is only available from ground inventory. Geostatistical techniques are used here to create an interpolated dataset, a 'map' of individual species distribution, from known sample information. In a previous study, we found that indicator kriging and sequential gaussian conditional simulation (sgCS) were promising tools for estimating sugar maple distribution from the USDA National Forest Inventory and Analysis (FIA) data. The techniques provided an estimate of species occurrence and a measure of uncertainty associated with that estimate, while retaining much of the local variability present in the sample data. In this study, these techniques are applied to 9 additional species in Pennsylvania. Four output datasets are available for each species--the probability of species occurrence, an estimate of its relative abundance, and a plus and minus level of uncertainty associated with that estimate. The datasets, used in conjunction with one another, provide the user with considerable flexibility in setting up the balance of errors of omission and commission that best suit the analysis under consideration. Similarities and differences between the species are identified and discussed as to their possible effect on the final estimates. Examples of how the datasets can be used are also presented. Indicator kriging and sgCS, used in conjunction with FIA sample data, provide a relatively straightforward technique to describe species occurrence and relative density across a state.

---

## INTRODUCTION

Data describing forest composition--so desired as a basic data source for many aspects of ecosystem analyses, models, and management--are generally unavailable and/or is stored in fixed forest-type categories. But forest communities are often not well characterized by the discrete categories imposed by forest cover type divisions. Inherent in each forest type category is an entire continuum (usually multi-dimensional) of different species and their relative importance.

In a previous study, we compared the geostatistical techniques available to interpolate FIA sample data to create a 'map' of tree species distribution.<sup>2</sup> The

---

<sup>1</sup>Northeastern Forest Experiment Station, USDA Forest Service, 5 Radnor Corporate Center, Radnor, PA 19087-4585.

tools of ordinary kriging, multigaussian kriging, indicator kriging, and sequential gaussian conditional simulation (sgCS) were used to estimate the occurrence and distribution of sugar maple in Pennsylvania. After considering the phenomenon being examined, the sample data being used, and the kind(s) of output desired in this study, we decided that indicator kriging sgCS proved to be the best interpolation tools for:

- a) providing an estimate of sugar maple occurrence,
- b) providing an estimate of sugar maple 'importance' in terms of %ba/acre,
- c) providing a measure of uncertainty associated with the two estimates,
- d) maintaining local variability,
- e) maintaining the characteristics of the original sample data, and
- f) handling sample data with highly skewed distributions.

a) An estimate of each species' presence or absence was provided by indicator kriging. An indicator transform divides the data into two classes--either above or below a designated cutoff value; in this study 0%ba/acre indicating presence or absence. Indicator kriging calculates for each cell an estimate of the probability that it falls above or below the cutoff value. The output dataset thus indicated the probability that sugar maple occurred at each location.

b) The second piece of information desired was an estimate of the relative amount of sugar maple at that location--i.e., whether the species represented a minor, moderate, or a major component of the total ba/acre on the plot at that location. Sequential gaussian conditional simulation determines multiple estimates for each cell. All are equally probable, and yet alternative realizations of the data determined from multiple simulation runs. From this set of estimates, an entire distribution can be built for each cell, representing the range of possible values. A summary statistic such as the mean or median of this distribution can then be chosen and used as the modeled 'estimate' of %ba/acre for that cell.<sup>3</sup>

c) A level of uncertainty is always associated with any estimate. Knowing how much uncertainty exists will help the user identify whether that uncertainty is acceptable for a specific task and how the data can be used. Correspondingly, knowing how much uncertainty exists will help the producer identify areas in which additional sampling would most improve the estimates. For estimates of species' presence/absence, indicator kriging provides this information in terms of a probability. For estimates of %ba/acre, summary statistics such as standard deviation or inter-quartile range were calculated from the distribution of simulated values to describe the variation associated with the %ba/acre estimate for each cell.

---

<sup>2</sup> Hershey, R. Riemann, M.A. Ramirez, and D.A. Drake. Using Geostatistical techniques to estimate the distribution and relative density of individual tree species in Pennsylvania. Unpublished report on file at USDA Forest Service, Northeastern Forest Experiment Station, Forest Inventory and Analysis Unit, Radnor, PA.

<sup>3</sup> The methods are described in Rossi et al. 1993 and Isaaks and Srivastava 1989; the analysis was performed using GSLIB routines (Deutsch and Journel 1992) with some additional routines written by R.E. Rossi.

d) Tree species in Pennsylvania exhibit a high level of local variation as a result of natural environmental factors and land use histories. At the intensity of sampling present in the FIA sample data, much of this local variation cannot be modeled and effectively predicted in the interpolation process, but instead appears as variation that is unexplained by neighboring plots. However, such local variability is an important characteristic of the distribution of a species. Thus, we did not want this local variability to become hidden behind a regional average of the resource, but to remain as apparent and accessible to the user as possible in the final estimated dataset(s). Sequential gaussian conditional simulation was the most effective of the interpolation methods at maintaining local variation.

e) One feature of a well-designed sampling scheme is that it is sensitive to and can report, with an acceptable level of error, the characteristics of the phenomena of interest. In this ideal situation, the characteristics of the sample data represent reasonably well those characteristics of the phenomena itself. Every estimation technique honors and maintains different aspects of the original data. The specific goals of the interpolation task at hand will determine the priorities, but in general the more characteristics of the sample data that are preserved in the estimated dataset, the more desirable the dataset. Sequential gaussian conditional simulation again did the best job of maintaining both the univariate and bivariate characteristics of the sample data.

f) As is true with many plant and animal populations, tree species have population distributions that are distinctly skewed toward younger individuals--more small trees than large mature ones. In addition, Pennsylvania, like most of the northeastern states, contains primarily mixed forests. Individual species rarely occur in pure stands. The 10 species examined in both studies included 8 of the most common by volume in Pennsylvania, and yet more than 50% of the time, when a species occurred on a plot it occurred as only a minor component (here defined as making up less than 20% of the total ba/acre on that plot). Both factors are combined in the %ba/acre 'relative importance' value, resulting in a highly skewed frequency distribution. Such extreme characteristics in the sample data can cause difficulties and biases when used with some of the interpolation methods that depend on assumptions about the normality of the distribution of the sample data (Isaaks and Srivastava 1989). One particular advantage of indicator kriging is that it makes no assumptions about the distribution of the data. Sequential gaussian conditional simulation, on the other hand, does assume that the data are normally distributed and stationary, and must be used more carefully. The sgCS routine used here, from Deutsch and Journel (1992), performs a 1-1, invertible normal-score transform on the data before running the simulation. In addition, however, the data also should be checked for binormality and a decision made as to whether to assume multivariate normality before the results of conditional simulation are accepted (Rossi et al. 1993).

The data exploration techniques used also proved invaluable to understanding the spatial characteristics of the species/variable being examined. Techniques included univariate analysis, variograms and other spatial dependence analyses,

and calculating local statistics. The resulting information was critical not only in determining what interpolation methods were most suitable and for checking the sample data for errors, but also for understanding the characteristics of the sample data and thus the phenomena being investigated. Geostatistical techniques offered ways to explore, organize, and summarize spatial patterns in the data that can provide clues to the variation and spatial behavior of the individual species under investigation.

### **Applying the techniques to 10 species in Pennsylvania**

Because of the promising results from using geostatistical techniques for estimating the distribution of sugar maple from FIA data in the previous study, the same geostatistical methods were applied to nine additional species: red oak, white oak, chestnut oak, black oak, hemlock, red maple, beech, white pine, and yellow birch. This list includes 8 of the top 10 most abundant species in Pennsylvania by volume, and two species (yellow birch and white pine) that are much less common (Alerich 1993).

Tree species distribution is affected by many factors, including both environmental conditions and direct human influence through harvesting and other land use histories. As a result of being differentially affected by all of these factors, each species will exhibit different patterns and scales of spatial distribution. Some of these factors occur at scales much smaller than the sampling intensity of the FIA data, and some occur over larger areas, representing broad-scale variation in the species distribution. In the previous study, it was found that a substantial amount of variation in sugar maple distribution was resolved at the sampling scale used for the FIA plots. This spatial dependence could, therefore, be modeled and used to support estimates of species occurrence and relative 'importance' (%ba/acre). The goal of this study is to examine to what extent this is true for the other species. More specifically, the objectives of this study are:

- a) if spatial dependence is exhibited at this intensity/scale of sampling,
- b) the resulting spatial distribution for each species, and how that compares to our current understanding,
- c) how the species differ from one another in terms of spatial dependence and distribution, and how that affects our ability to estimate them, and
- d) how to use the resulting estimated datasets.

### **DATA**

The sample data were collected by the Northeastern Forest Experiment Station's Forest Inventory and Analysis (FIA) unit. Basal area--the summed cross-sectional area at breast height--is calculated for all live trees 1.0 inches DBH or larger on the plot (Hansen et al. 1992). The data were for individual tree species, by basal area (ba) per acre as a proportion of the total basal area (% ba/acre). The data were accessed from individual tree records in the USFS Eastwide tree-level database and summarized as %ba/acre for each species by plot. In Pennsylvania, there were a total of 5,100 plots. Nonforested plots and those with total ba/acre equal to zero (due to missing data) were removed--leaving only 2,905 plots.

## METHODS

Each species was examined entirely independently. As in the previous study, the data for each species was organized, summarized, and explored using univariate statistics, measures of spatial dependence (variogram, covariance, and correlogram), and spatial distribution of local statistics across the state. All species were similar in many of their basic characteristics to each other and to the previously investigated species, sugar maple. Each exhibited extremely skewed distributions, with more than 50% of the plots containing less than 1%ba/acre in every species except red maple and red oak. A variogram was calculated for both the raw sample data and for a 1-1, invertible normal-score transform of the sample data, using a lag distance of 500m and no directional component (anisotropy). In every instance, the variogram of the normal-scored data exhibited considerably more spatial dependence and structure than that of the raw data (Figure 1), revealing spatial characteristics that were hidden by the strong univariate characteristics of the data. As sgCS uses normal-scored data, it was the model fitted to the normal-scored variogram that was used in the conditional simulation. An indicator variogram also was calculated and modeled for use in the indicator kriging. In general, there was far less structure and less of the variation explained in the indicator variogram (32 to 57%) than in the normal-scored regular variogram (35% and 64 to 97%) (Table 1). To assess how areas of 'local' variability in the sample data changed across the state, the mean and standard deviation were calculated for each of the 23,400 3000 x 3000m cells, using a 15 x 15km area as the window defining the size of the 'local' area. All species exhibited a proportional effect, with areas of high mean corresponding with areas of high local standard deviation, indicating a lack of stationarity. Using normal-scored data seemed to largely eliminate this situation.

**Table 1. The percent variation explained by the spatial dependence in the variograms.**

Species	Indicator variogram	Normal-scores variogram
Beech	46	76
Black oak	37	75
Chestnut oak	44	82
Hemlock	57	66
Red maple	32	35
Red oak	38	64
Sugar maple	32	97
White oak	50	79
White pine	37	84
Yellow birch	39	81

Indicator kriging and sgCS were run for each species, using models derived from the appropriate variograms. The estimation parameters of cell size (3,000m), search radius (10,000m), and minimum:maximum number of points used (1:16) were taken directly from the results of the previous study. In sgCS, 50 simulations were run for each species.

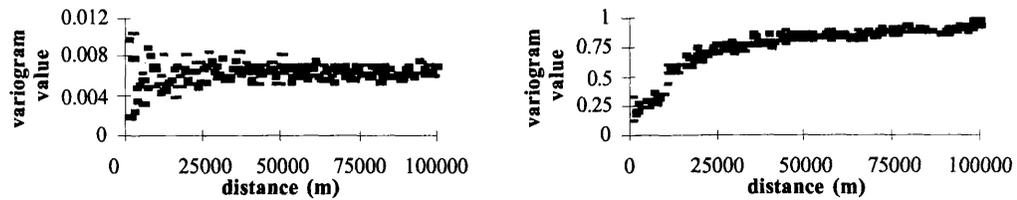


Figure 1. Spatial dependence as demonstrated by the variogram of the raw data (left side) and variogram of the normal-scored data (right side) for white pine.

## RESULTS

With the exception of red maple, all species examined demonstrated substantial spatial dependence in the variogram of normal-scored data, with 64 to 97% of the variation explained by the visible structure and capable of being modeled. In some species, much of that spatial dependence was contained in a very long-range trend of about 100,000m. White oak, black oak, chestnut oak, and beech all fell into this category. The rest of the species appeared to split the bulk of the explained spatial dependence over 2 ranges. For red oak, this was both a short- and medium-range pattern (12,000 and 40,000m); sugar maple a very short- and a long-range pattern (2,100 and 60,000m); and yellow birch a medium- and long-range pattern (19,000 and 80,000) (Figure 2). The spatial dependence exhibited in the indicator variograms was much less, ranging from only 32 to 57% of the variation explained. This is not ideal for interpolation and suggests the necessity for further refinement.

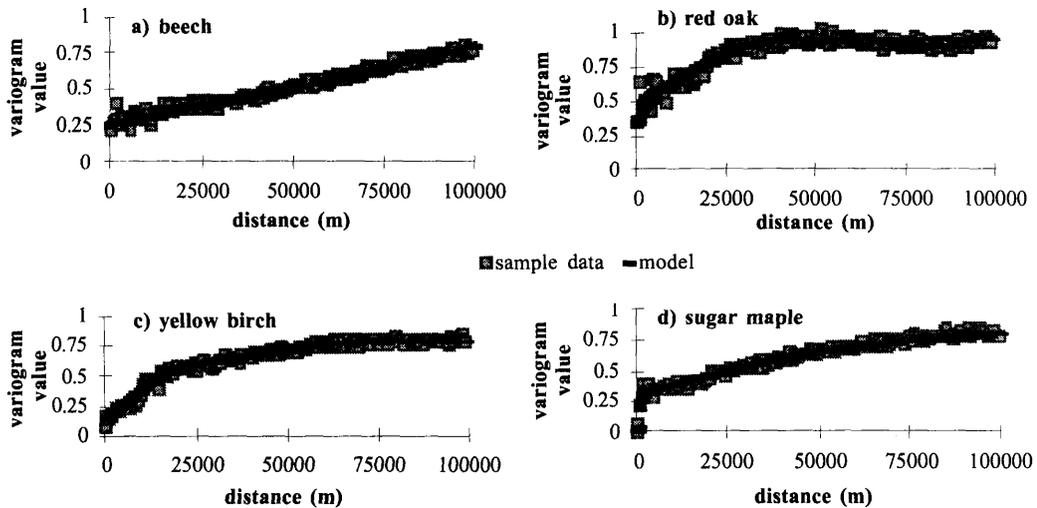
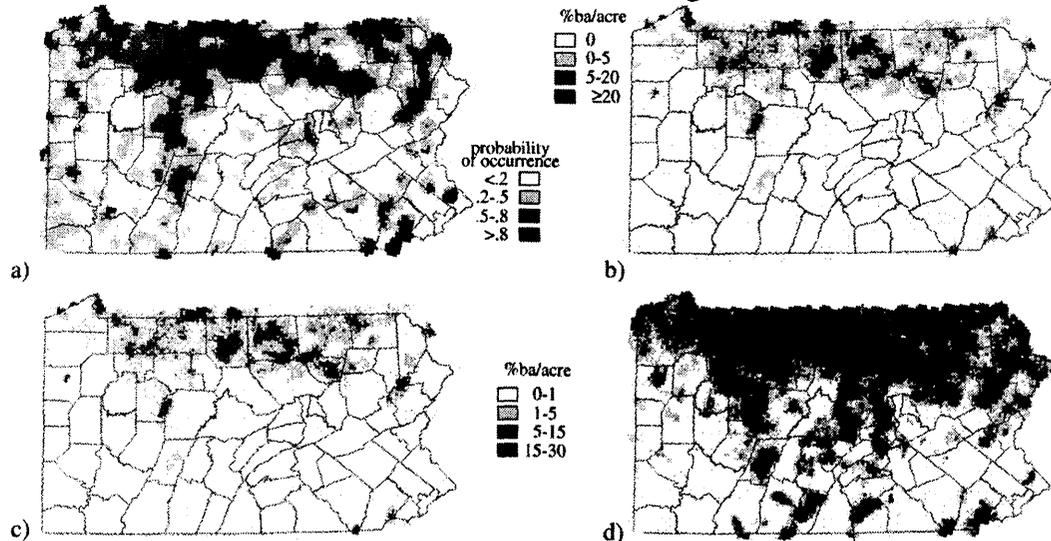


Figure 2a-d. Variograms from the normal-scored data for four of the species.

The results create several output datasets. Figure 3a-d shows the results for beech, using 4 datasets to represent the species. Part (a) shows the estimated probability of beech occurrence, as calculated from indicator kriging. Part (b) is the median value of the 50 sgCS simulations, representing the chosen estimate of beech relative ‘importance’ in %ba/acre. The uncertainty associated with that estimate (here chosen to be a percentile range capturing approximately 2/3 of the

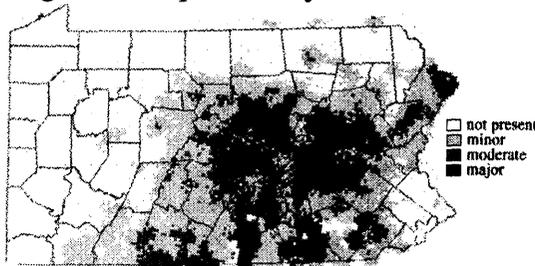
distribution) is described in (c) and (d). Part (c) expresses the minus variation, or that distance in %ba/acre values between the median and the bottom of that range (the 17th percentile), and part (d) expresses the plus (+) variation (83rd-50th percentiles). In every instance, the + variation is much greater.



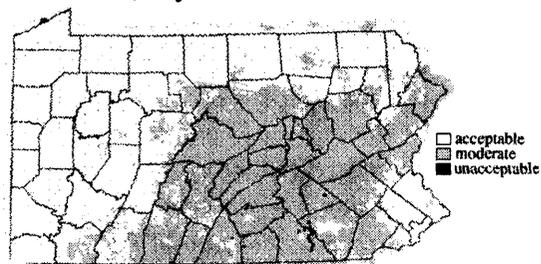
**Figure 3.** Four estimated datasets describing the distribution of beech in Pennsylvania: a) the estimated probability of occurrence using indicator kriging, b) the median values from 50 sequential gaussian conditional simulations, c) the minus variation (median-17th percentile), and d) the plus variation (83rd percentile-median) about the median estimate.

## DISCUSSION

The probability that a species occurs is a unique and useful dataset. Any probability can be used as the cutoff to create a species presence/absence map depending on the objectives of the task at hand. If, for example, a particular insect is known to live in chestnut oak forests and the objective is to limit the search to only those areas where there is a high probability of finding suitable conditions, we might set the cutoff at the probability level of  $\geq .8$ . If, however, we are most interested in not missing any areas where the insect occurred, we might set the probability level for forest much lower, say  $\geq .4$ .



**Figure 4.** Chestnut oak occurrence as a major (>40%ba/acre), moderate (20-40%), and minor component (<20%) of the total ba/acre. Derived from sgCS estimates using the 75th percentile.



**Figure 5.** The uncertainty associated with the estimates used in Figure 4, in classes of: acceptable ( $\leq \pm 10\%$ ba/acre), moderate ( $\pm 10-40\%$ ), and unacceptable ( $> \pm 40\%$ ). Derived from the sgCS estimates.

The distribution of densities at which a species occurs also can be of considerable interest. For example, Figure 4 is a plot produced from sgCS estimates illustrating where chestnut oak occurs as a major, moderate, and minor component. The three categories have been defined as 1 to 20%ba/acre, 20 to 40%ba/acre, and >40%ba/acre. The %ba/acre estimate in Figure 4 can be associated with a corresponding uncertainty dataset (Figure 5) in which the uncertainty classes are broken into acceptable, moderate, and unacceptable. Again, the dataset itself is unclassified, and the user defines these--all the building blocks are provided in the output from sgCS.

Alternatively, the same %ba/acre dataset(s) can be used to create a forest cover type 'map.' For this purpose, the summary statistic (whether mean, median, or another of the percentiles of each cell's simulated distribution) can be chosen specifically for the intended purpose in the same way the different levels of probability could be chosen when mapping species presence/absence from indicator kriging estimates. If the objective is to reduce the error of commission (i.e., classifying areas as sugar maple/beech (SM/B) in the estimated map that really are not SM/B), then using a percentile at the lower end of each cell's distribution would be more desirable. If, however, the objective is to reduce the error of omission (i.e., missing areas that do contain SM/B), then using a higher percentile from the distribution, such as 75%, would be more desirable.

The %ba/acre estimates of three species were further refined. Two species, white pine and yellow birch, nearly disappeared in the initial sgCS estimate even though they showed up as present in the probability of occurrence estimate from the indicator kriging. This was considered suspicious, so the dataset for each of these two species was divided into several different populations by region, and the process of variogram modeling and interpolation was repeated for each region. The regionalized variogram was substantially different from the average variogram, and when sgCS was performed using the locally tuned models, it revealed an entirely different and much more credible picture of white pine distribution in Pennsylvania. The third species, hemlock, was refined based primarily on the suspicion that small and large stand-size classes may have different spatial distribution patterns. When variograms were calculated separately for these two populations (using a cutoff of 45 years), they were indeed substantially different in shape, sill, nugget, and range.

## CONCLUSIONS

The estimated datasets output by indicator kriging and sgCS have the potential to be very useful. Every species examined, with the exception of red maple, exhibited substantial spatial dependence in the variograms of the normal-scored data, suggesting that there is considerable potential for the estimation of such datasets from FIA data. Each species exhibited some variety in spatial patterns and spatial dependence and may require different levels of additional fine-tuning, depending upon the objectives of the specific analysis and the time and expertise available.

### **Characteristics that make the datasets useful**

These techniques make explicit the uncertainties associated with an estimate in a form that can be incorporated when the data are used. This feature adds considerable utility and flexibility in the use of the resulting estimates, as the risk of errors of commission or omission can be specifically determined and manipulated to suit the current objectives.

Maintaining individual species information separately allows considerable flexibility in the use of species distribution data. Instead of being limited to previously defined fixed classes, forest cover types can be uniquely defined to capture more accurately the habitat required for a particular study. The potential also exists to use one or more of the species datasets as a decision layer in the interpretation of satellite imagery. The two datasets offer complementary information about the species composition that really exists on the ground.

The techniques used in this study are not extremely time-consuming nor difficult to process, and could be easily extended to additional species and states. There is a high level of variance associated with these estimates of %ba/acre--in many locations this variance can be as much as the estimate itself. However, the dataset nevertheless provides a very descriptive picture of species distribution at the state level. In comparison to previous depictions of current species distribution from FIA data by summarizing at the county level, this method provides a much more detailed picture of species occurrence and distribution. Although estimates could probably be improved and variances diminished by additional investigation into each species, the current estimates are informative and provide a useful basis from which to proceed.

### **Clues to refining the interpolation and improving the estimates**

As was observed with white pine, yellow birch, and hemlock, it may be possible to significantly improve the estimates of %ba/acre by refining the analysis. When sub populations of a species have a significantly different pattern of spatial distribution, treating the populations separately in the interpolation will improve the final estimates. These populations may be described by regional land features or by some other defining characteristic (e.g., stand age for hemlock). The results suggest several clues to determine when this additional effort is necessary. First, when results of indicator kriging showed a species as present, but sgCS did not. Second, when the spatial dependence in the variogram is noticeably less than expected. This may suggest that there are different populations being lumped together that should be separated. For white pine, dividing the state into several broad geographic regions made a significant difference in the calculated variogram and thus in the final estimates. Another important clue is previous knowledge about the species that different ecological regions may have caused distinct spatial distribution patterns among tree species, or that different size class or age populations may have different spatial distribution patterns over the landscape. Hemlock is an example of the latter. As a result of past management practices that involved heavy harvesting of large hemlock for the tanning industry, today there are often relics of large individuals among a relatively wider

distribution of smaller, younger trees that have grown up in the interim (Hough and Forbes 1943, Powell and Considine 1982).

These datasets of individual species distribution do not contain any of the fine-scale forest/nonforest detail. If such information is desired, more detailed datasets describing the forest/nonforest land cover in Pennsylvania would have to be derived from a more intense point sample or the continuous but averaged data available from satellite imagery. Such detailed datasets could be used as a 'mask' and overlaid on any of the datasets of species distribution.

There are more possibilities for applying geostatistical techniques than have been investigated here. For example, some species may exhibit some correlation with a particular soil, climate, topography, or reflectance data from satellite imagery. Indicator kriging and sgCS, in particular, allow the incorporation of such ancillary, 'soft' information to contribute to the estimation process.

## REFERENCES

- Alerich, C.L. 1993. Forest Statistics for Pennsylvania--1978 and 1989. Resource Bulletin NE-126. USDA Forest Service, Northeastern Forest Experiment Station. Radnor, PA. 244p.
- Deutsch, C.V. and A.G. Journel. 1992. GSLIB: Geostatistical Software Library and User's Guide. Oxford University Press, New York.
- Hansen M.H., T. Frieswyk, J.F. Glover, and J.F. Kelly. 1992. The Eastwide Forest Inventory Data Base: Users Manual. General Technical Report NC-151. USDA Forest Service, North Central Experiment Station. St. Paul, MN.
- Hough, A.F. and R.D. Forbes. 1943. The Ecology and silvics of forests in the High Plateaus of Pennsylvania. *Ecological Monographs*. 13:299-320.
- Isaaks, E.H. and R.M. Srivastava. 1989. *An Introduction to Applied Geostatistics*. Oxford University Press, New York.
- Powell, D.S. and T.J. Considine. 1982. An analysis of Pennsylvania's forest resources. Resource Bulletin NE-69. USDA Forest Service, Northeastern Forest Experiment Station. Broomall, PA. 97p.
- Rossi, R.E., P.W. Borth, and J.J. Tollefson. 1993. Stochastic simulation for characterizing ecological spatial patterns and appraising risk. *Ecological Applications*. 3(4):719-735.

## BIOGRAPHICAL SKETCH

Rachel Riemann Hershey is a forester/geographer with the Forest Inventory and Analysis Unit, Northeastern Forest Experiment Station. She received a B.A. in ecology from Middlebury College, an M.S. in forestry from the University of NH and an M.Phil. in geography from the London School of Economics.

# A Calibration-Based Model for Correcting Area Estimates From Coarse Resolution Land Cover Data

Aaron Moody\*

---

**Abstract.**—A two stage modeling strategy significantly improves area estimates by correcting coarse-resolution measurements of class proportions. *Stage I* models use measurements of scale-invariant landscape spatial properties to estimate the slope and intercept of proportion transition relationships. A *stage II* model uses a regression estimator to predict true class proportions based on measured coarse-scale proportions, and the slope and intercept estimates from the *stage I* models. Model development and testing on a *calibration* site is followed by testing and inversion for a *validation* site. Inversion involves using spatial variables measured at the *coarse* scale as input to the *stage I* models. A probabilistic sampling strategy allows statistical assessment of the models and results.

---

## INTRODUCTION

Image spatial resolution will influence land-cover area estimates made from classified remotely sensed data (Mayaux and Lambin 1995; Moody and Woodcock 1994). Models of these resolution effects can lead to improved area estimates derived from coarse resolution remote sensing. One might consider two approaches for modeling this scale-dependent areal bias. Mixture models, incorporated into the classification process, can estimate the subpixel composition of pixels if pure-class spectra are known (Adams et al. 1986). Alternatively, calibrated models, applied in a post-classification mode, can improve area estimates from coarse-resolution data if the relationship between "true" and measured proportions can be modeled (Mayaux and Lambin 1995; Kalkhan et al. 1995).

This paper describes the development and evaluation of three linked, statistical models that provide post-classification correction of area estimates. The first two models estimate the slope and intercept of a line characterizing the relationship between true and coarse-scale proportions. In this case, "true" proportions refer to proportions measured at 30 m, and coarse-scale proportions refer to measurements at 1020 m. The slope and intercept are modeled based on a small set of relatively scale-invariant, measurable spatial properties of the landscape. The third model uses these estimated slopes and intercepts to predict 30 m proportions based on measurements of proportions at 1020 m.

---

\*Assistant Professor, Department of Geography, University of North Carolina, Chapel Hill, NC. 27599

## BACKGROUND

Two basic spatial effects contribute to biased area estimates. A *first order* effect is the tendency of large classes to increasingly dominate the landscape when measured at increasingly coarse scales. Accordingly, small classes tend to diminish in size. Figure 1 illustrates these patterns for all the data used in this study. *Second order* effects refer to modulations of these basic patterns due to specific landscape spatial organization (Moody and Woodcock 1995). These effects result in the scatter about the smoothed fit notable in Figure 1d. In either case, scale-dependent changes in the apparent area of classes result from class membership transitions between scales. This effect can be thus characterized in terms of proportion transition lines relating true and coarse resolution proportions. The slope of such a line can summarize the rate of transition.

If transition rates depend partly on landscape spatial organization, it is sensible to try to model them using measures of spatial pattern. A variety of spatial measures exist, and many reviews and summaries are present in vegetation analysis and landscape ecology literature (Legendre and Fortin 1989; Cullinan and Thomas 1992). In this research, landscape spatial properties are used to model the slopes and intercepts of the proportion transition lines for a large set of sampling units in independent calibration and validation sites. These are the *stage I* models. *Model inversion*, will require that the landscape pattern measures are relatively scale-invariant, or resistant to resolution.

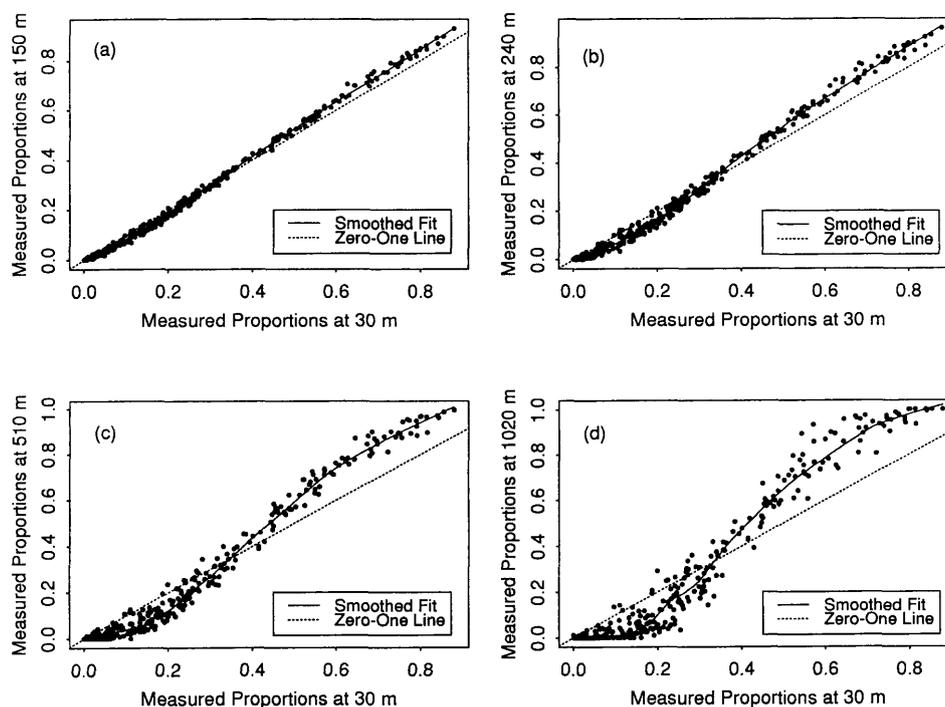


Figure 1.—Relationship between proportions at 30 m and at 4 coarser scales.

There are several calibration-based methods for improving coarse resolution area estimates (Czaplewski and Catts 1992; Kalkhan et al. 1995). The model used here is a form of "classical" model in which known but incorrect values are estimated using unknown correct values (Brown 1982). For example, if  $P_0$  represents true proportions and  $P_r$  represents measured proportions at some coarse spatial resolution  $r$ , then:

$$P_r = \beta_0 + \beta_1 \cdot P_0 + error \quad (1)$$

where  $\beta_0$  and  $\beta_1$  are the intercept and slope of the proportion transition line that relates true proportions to proportions measured at resolution  $r$ . Inverting this simple model provides:

$$\hat{P}_0 = \frac{P_r - \hat{\beta}_0}{\hat{\beta}_1} \quad (2)$$

which is the *stage II* model used in this paper.

A five part strategy is employed as follows: *a*) identify a set of scale invariant spatial measures; *b*) use a subset of these measures (calculated at 30 m) ( $X_{30}$ ) to develop *stage I* models for predicting the slope ( $\hat{\beta}_1$ ) and intercept ( $\hat{\beta}_0$ ) of the proportion transition lines for the *calibration* site; *c*) predict 30 m proportions ( $\hat{P}_0$ ) for the calibration site by applying the *stage II* model in Eq. 2 using the measured 1020 m proportions ( $P_r$ ), and the slopes ( $\hat{\beta}_1$ ) and intercepts ( $\hat{\beta}_0$ ) estimated from the *stage I* models; *d*) using the *stage I* models developed on the calibration site, repeat step *c* and evaluate the procedure when applied to the data from the *validation* site; *e*) invert and evaluate the procedure by running the *stage I* models using the spatial variables as measured at 1020 m ( $X_{1020}$ ), and supply the results to the *stage II* model to predict 30 m proportions based *only* on information measured at 1020 m for the validation site.

## METHODS and RESULTS

The Plumas and Stanislaus National Forests are used as calibration and validation sites, respectively. Landsat Thematic Mapper data have been classified to produce maps of general land-cover categories for each site. Classes include *barren*, *brush*, *hardwood*, *water*, and *conifer*. Although the two sites have similar characteristics, they are spatially separated by roughly 2° of latitude.

The data for each site are aggregated to 1020 m resolution using a plurality-based aggregation procedure. This involves coding each grid cell in a 1020 m sampling grid with the most frequently occurring subgrid-cell class.

A set of randomly located 238x238 pixel subregions serve as the sampling units for the analyses. Each unit contains 56,644 30 m pixels and 49 1020 m pixels. The Plumas contains fifty sampling units for model development and initial testing. The Stanislaus contains thirty-five units for model validation and model inversion. The number of units represent 30% of all *possible* such units from each site. Within each sampling unit the following measurements are collected: proportions at 30 m for each class, proportions at 1020 m for each class, a set of spatial measures at 30 m, and the same set of spatial measures at 1020 m. The

ultimate goal is to estimate 30 m proportions by supplying 1020 m area measurements and slope and intercept coefficients to the *stage II* model (Eq. 2). An intermediate goal is to estimate the proper slopes and intercepts using a multiple regression model with a parsimonious set of spatial measures as the independent variables. The slopes and intercepts of the proportion transition lines are the dependent variables in these *stage I* models.

A variety of spatial measures are determined within each unit using the *r.le* software (Baker and Cai 1992). Five of these demonstrate scale-invariance as determined by examining the simple correlations between each variable and itself at the two different scales. Of these five, three prove significant in modeling both the slope and the intercept of the proportion transitions as determined within each sampling unit. An additional variable (*c* below) that does not have the scale invariance property is included because it characterizes an important landscape characteristic not previously included in the model. The four variables used are: **maximum class size (*mx*)**, **inverse Simpson's index (*s*<sup>-1</sup>)**, **contagion (*c*)**, and **entropy (*ent*)**. Expressions for the latter 3 variables are:

$$s^{-1} = \frac{1}{\sum_{i=1}^k P_i^2} \quad (3)$$

$$c = 2 \cdot \ln(k) - ent \quad (4)$$

$$ent = -\sum_{i=1}^k \sum_{j=1}^k P_{ij} \cdot \ln(P_{ij}) \quad (5)$$

where *k* is the number of classes present, *P<sub>i</sub>* is the proportion of class *i* in the sampling unit, and *P<sub>ij</sub>* are elements of a *k* × *k* co-occurrence matrix and represent adjacency probabilities between classes *i* and *j*. Maximum class size refers to the proportion of the largest class in the sampling unit. Simpson's index indicates the probability of randomly selecting two pixels of the same attribute. Contagion measures the degree of clumping in the landscape. Entropy is maximized when all pixels of a given class are as far away from one another as possible. Table 1 shows the cross-scale correlation matrix for this set of variables.

Table 2 presents regression summaries for the two *stage I* models. The independent variables of are the spatial variables measured at 30 m as described above. The dependent variables are *a*) the intercepts and *b*) the slopes of the proportion transition lines as determined using a linear least squares fit between the 30 and 1020 m proportions for the classes existent within each sampling unit.

Table 1.—Cross-scale correlations (scale-invariance) of independent variables.

	<i>mx</i> <sub>1020</sub>	<i>s</i> <sup>-1</sup> <sub>1020</sub>	<i>c</i> <sub>1020</sub>	<i>ent</i> <sub>1020</sub>
<i>mx</i> <sub>30</sub>	0.91	--	--	--
<i>s</i> <sup>-1</sup> <sub>30</sub>	--	0.69	--	--
<i>c</i> <sub>30</sub>	--	--	-0.07	--
<i>ent</i> <sub>30</sub>	--	--	--	0.81

Table 2.—Stage I models. Slope model  $R^2_{adj}=0.63$ . Intercept model  $R^2_{adj}=0.70$ .

Slope Model	Coefficient	Standard Error	t-value	P >  t
$\beta_0$	-5.77	1.05	-5.49	0.00
$mx_{30}$	4.95	0.84	5.92	0.00
$s^{-1}_{30}$	-0.74	0.22	-3.33	0.002
$c_{30}$	-0.679	0.17	-4.06	0.0002
$ent_{30}$	2.83	0.49	5.83	0.00
<b>Intercept Model</b>				
$\beta_0$	0.0042	0.093	0.045	0.96
$mx_{30}$	-0.30	0.074	-4.054	0.0002
$s^{-1}_{30}$	0.067	0.020	3.3904	0.0015
$c_{30}$	0.113	0.015	7.68	0.00
$ent_{30}$	-0.145	0.043	-3.39	0.0015

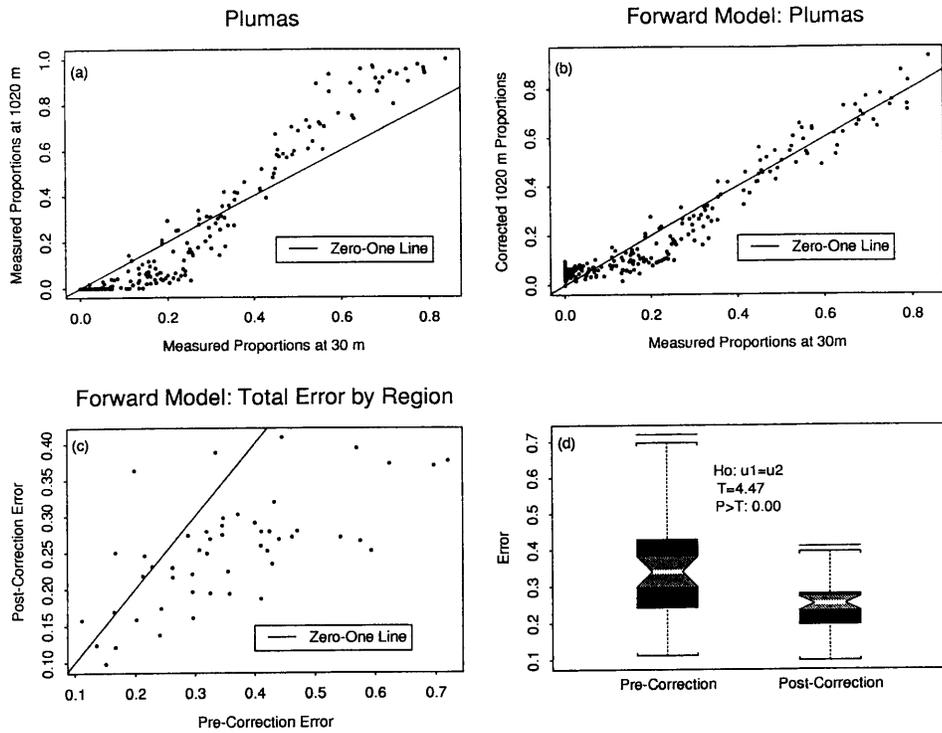
The *stage I* models described in Table 2 estimate the  $\hat{\beta}_0$  and  $\hat{\beta}_1$  coefficients necessary for employing the *stage II* model (Eq. 2). This sequential modeling process is conducted three times. First, the *stage I* models are developed using the data from the Plumas (calibration site). The predicted values from these models are then used in Eq. 2 to test the overall modeling process for the calibration data. Second, the *stage I* models are applied in a predictive mode using the independent variables as measured at 30 m from the Stanislaus (validation site). Again predicted  $\hat{\beta}_0$  and  $\hat{\beta}_1$  values supply the coefficients to run Eq. 2 and estimate 30 m proportions for the Stanislaus. Third, *stage I* models are applied using the variables measured at 1020 m from the Stanislaus and the results are again used to estimate 30 m proportions. The first two cases are *forward* models in the sense that they require high resolution information to perform the correction. The third is an *inverted* model, because it relies only coarse resolution data. The results from these three tests are presented in Figures 2, 3, and 4.

## DISCUSSION and CONCLUSIONS

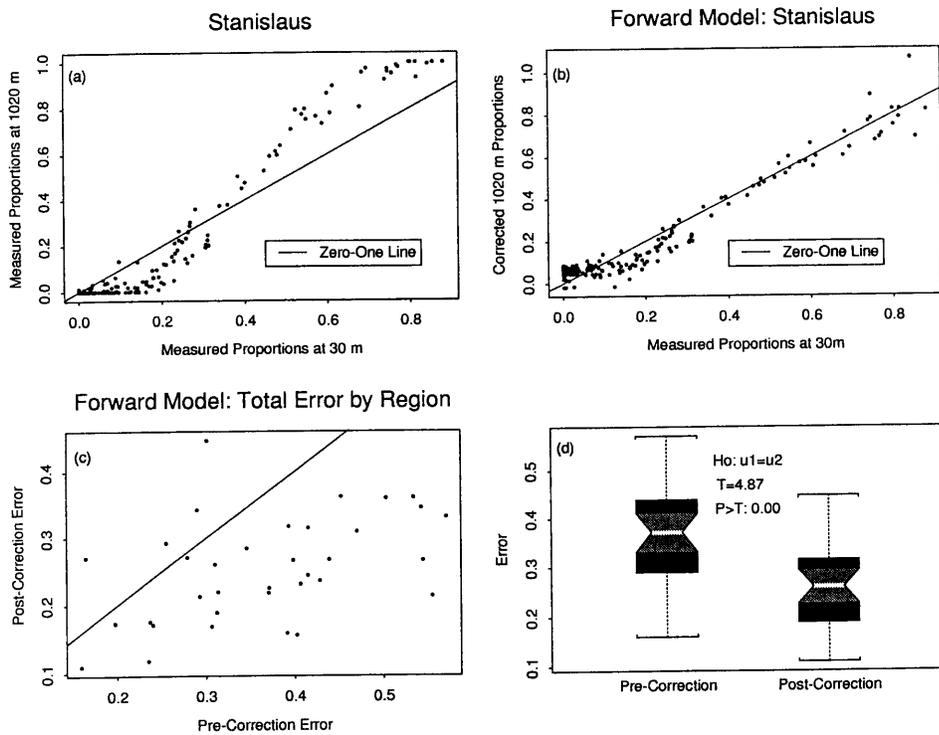
As seen in Figure 1, Figures 2a, 3a, and 4a also illustrate the basic scaling effects for class proportions. Note that at 1020 m, the greatest underestimations occur for intermediate-small classes, and the greatest overestimations occur for intermediate-large classes. Very large, very small, and moderate sized classes (at the cross-over point, around 30%) are all reasonably estimated at 1020 m.

The goal of the two stage modeling procedure is to pull the coarse-resolution area estimates closer to the zero-one line. Figures 2b and 3b demonstrate that the model improves area estimates for both the calibration and the validation sites when operated in the forward mode. The correction procedure performs best for large classes. For small classes a notable dip (albeit reduced) below the zero-one line still occurs. An intercept effect is also evidenced by the vertical alignment of estimates above the zero value of the x-axis. Results from the inverted model (Figure 4b) also show general improvement, although considerable scatter occurs for very large classes (note the 2 outliers).

Once corrected values are derived, it is possible to tabulate the total absolute error within each sampling unit. For a given sampling unit  $g$ , the total error is:



**Figure 2.—Forward model results for the calibration site.**



**Figure 3.—Forward model results for the validation site.**

$$error_g = \sum_{i=1}^k |P_{i,1020} - P_{i,30}|$$

where  $P_{i,1020}$  and  $P_{i,30}$  are the proportions for class  $i$  at 1020 and 30 m, respectively. By calculating these values for both pre- and postcorrection data, the results can readily be compared. Figures 2c and 3c show the relationship between pre- and postcorrection error for the calibration and validation sites using the forward model. For any point falling below the zero-one line, the total error is reduced due to the correction procedure. At both sites, the total error for roughly 90% of the sampling units is either reduced or unchanged after correction. Conversely, the error for roughly 10% of the units is increased. For the inverted model (Figure 4c) 80% of the regions are either improved or unchanged after correction with two positive outliers (regions 21 and 25).

Figures 2d, 3d, and 4d are boxplots comparing the distributions of the pre- and postcorrection total error values. In all cases, the correction procedure results in a significant reduction in error. For the inverted model the test was performed after removing the outliers. In each case, however, the  $T$ -tests are suspect due to unequal variance.

Several interesting questions deserve continued attention. How do the spatial measures used govern the proportion transitions? Will the scale-invariance property transfer to other landscape types?

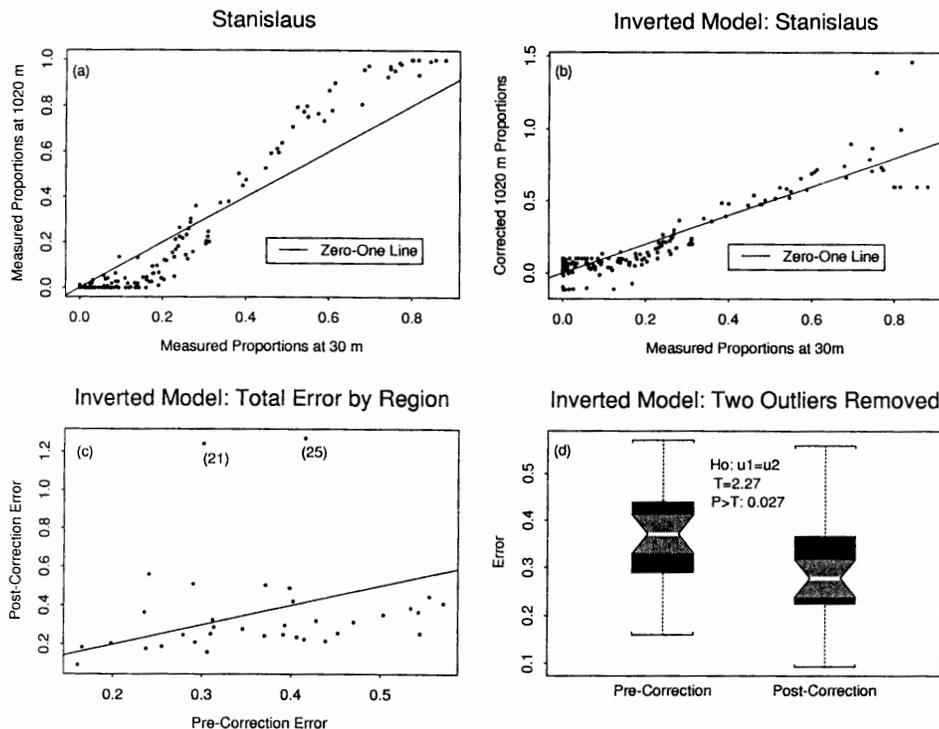


Figure 4.—Inverted model results for the validation site.

Might other scale-invariant measures better predict proportion transitions? Is the general procedure extensible across landscape types? What is the effect of constraining the intercept to zero? What are the sensitivities of the cross-over point seen in Figure 1? Resolving these and other issues will help formalize a body of understanding of how class proportions scale. This understanding hopefully will lead to improved land-cover area estimates at local, regional and global scales.

## REFERENCES

- Adams, J. B., Smith, M. O., and Johnson, P. E. 1986. Spectral mixture modeling: A new analysis of rock and soil types at the Viking Lander 1 site. *J. Geophys. Res.* 91(B8):8098-8112.
- Baker, W. L. and Cai, Y. 1992. The r.le programs for multiscale analysis of landscape structure using the GRASS geographical information system. *Landscape Ecology* 7(4):291-302.
- Brown, P. J. 1982. Multivariate calibration. *J. Royal Statistical Soc.* 3:287-321.
- Cullinan, V. I. and Thomas, J. M. 1992. A comparison of quantitative methods for examining landscape pattern and scale. *Landscape Ecology* 7(3):211-227.
- Czaplewski, R. L. and Catts, G. P. 1992. Calibration of remotely sensed proportion or area estimates for misclassification error. *Remote Sens. Environ.* 39:29-43.
- Kalkhan, M. A., Reich, R. M., and Czaplewski, R. L. 1995. Evaluation of statistical properties of the inverse estimator for remotely sensed areal estimates using simple random sampling. *Proc. Amer. Soc. Photogramm. and Remote Sensing Conf.*, 27 Feb. - 2 Mar. 1995, Charlotte, NC, 3:258-270.
- Legendre, P. and Fortin, M-J. 1989. Spatial pattern and ecological analysis. *Vegetatio* 80:107-138.
- Mayaux P., and Lambin, E. F. 1995. Estimation of tropical forest area from coarse spatial resolution data: A two step correction function for proportional errors. *Remote Sens. Environ.* 53:1-16.
- Moody, A. and Woodcock, C. E. 1995. The influence of scale and the spatial characteristics of landscapes on land-cover mapping using remote sensing. *Landscape Ecology* 10(6):363-379.
- Moody, A. and Woodcock, C. E. 1994. Scale-dependent errors in the estimation of land-cover proportions —Implications for global land-cover datasets. *Photogramm. Eng. Remote Sens.* 60(5):585-594.

## BIOGRAPHICAL SKETCH

Aaron Moody is a geographer at the University of North Carolina, Chapel Hill with a specialization in remote sensing of vegetation. He holds a Ph.D. from Boston University and an M.A. from the University of California at Santa Barbara, both in geography.

# Geostatistical Analysis of Multi-Spatial Resolution Imaging Spectrometer Data for Characterizing Forest Ecosystems

Paul M. Treitz<sup>1</sup> and Philip J. Howarth<sup>2</sup>

---

**Abstract.**—Multi-spatial resolution Compact Airborne Spectrographic Imager (CASI) reflectance data are analyzed using geostatistical techniques. Semivariograms are derived from visible and near-infrared data collected at two altitudes for selected forest ecosystems representative of the boreal landscape for northwestern Ontario, Canada. These analyses reveal that (i) mean ranges differ for ecosystems of different complexities; (ii) mean ranges for the near-infrared data are greater than the corresponding ranges derived for the visible data; (iii) mean ranges are greater for the medium-altitude data than for the low-altitude data; (iv) the height of the semivariogram (sill) increases for stands demonstrating complex layering, particularly for the near-infrared data; and (v) the sills are typically lower for the medium-altitude data. These results indicate that for optimal discrimination of forest ecosystems, multi-spatial resolution remote sensing data and spatial processing techniques offer potential for classification and mapping of forest ecosystems.

---

## INTRODUCTION

The development of the Forest Ecosystem Classification (FEC) for northwestern Ontario, Canada was spawned by the need to produce an ecologically based forest classification which incorporates both physiographic and biotic elements of the forest ecosystem. These detailed hierarchical classifications are designed for field-level mapping and are therefore difficult to implement for large tracts of forest that are characteristic of much of northwestern Ontario. Remote sensing data and digital image analysis techniques offer potential for assisting in the analysis of large tracts of forest for identification of relevant ecosystem classes, particularly in the context of an hierarchical classification scheme.

However, it is difficult to identify appropriate spatial resolutions for remote sensing data acquisition that will provide the most suitable level of information for extracting forest ecosystem parameters for a variety of environments. In this paper, the results of geostatistical analyses of multi-resolution Compact Airborne Spectrographic Imager (CASI) reflectance data for selected forest ecosystems in northwestern Ontario, Canada are presented. Using these types of analyses for various forest ecosystems at different spatial resolutions, it is possible to gain insight into the impact of spatial structure on information extraction from these data.

---

<sup>1</sup> Lecturer, Department of Geography, York University, North York, Ontario, Canada

<sup>2</sup> Professor, Department of Geography, University of Waterloo, Waterloo, Ontario, Canada

## GEOSTATISTICAL ANALYSIS OF REMOTE SENSING DATA

Jupp *et al.* (1988) describe many remote sensing scenes as spatial arrangements of two- or three- dimensional objects superimposed on a uniform background. In this scenario, they consider a discrete object scene model as an appropriate abstraction of the scene. In this model, one or more classes of objects can be described by a unique set of properties or parameters. This abstraction is appropriate for certain applications of remote sensing and digital image analysis. However, within a forest ecosystem context, not only does the canopy structure vary with ecosystem type, but the background can also vary significantly with respect to composition and structure.

The brightness values of a remotely sensed image are a function of spatial position during image acquisition. In the realm of geostatistics, these brightness values can therefore be considered as values of a “regionalized variable” (Matheron, 1963). To apply regionalized variable theory to the analysis of remotely sensed data, adoption of a stochastic view of the landscape and its spatial structure is required (Jupp *et al.*, 1988). This is a logical assumption since underlying processes and properties of the landscape will produce many similar scenes. In fact, this is intrinsic to the classification of ecosystems. It is assumed that similar ecosystems will arise from similar environmental conditions and processes linked closely to landscape. In this study, semivariogram analysis is used to examine the spatial structure of high resolution CASI data collected for natural forest ecosystems, characteristic of northwestern Ontario.

### Semivariogram Analysis

The semivariogram, which originates from the theory of regionalized variables developed by Matheron (1963), is used to measure the spatial dependence of neighboring observations for any continuously varying phenomenon. Hence, it is a technique that can be applied to spectral data, a variable for which position in space and time is known. In this manner, spatial variation in images can be examined in relation to ground scene and sensor parameters (Woodcock *et al.*, 1988).

The semivariogram plots semivariance against spatial separation along a given relative orientation and provides a concise and unbiased depiction of the scale and pattern of spatial variability (Curran, 1988). The semivariogram or  $\gamma(h)$  is calculated as:

$$\gamma(h) = \frac{1}{2(n-h)} \sum_{i=1}^{n-h} [Z(x_i) - Z(x_i + h)]^2$$

where  $h$  is the lag (or distance) over which  $\gamma$  (semivariance) is measured,  $n$  is the number of observations used in the estimate of  $\gamma(h)$ , and  $Z$  is the value of the variable of interest at spatial position  $x_i$ . The value  $Z(x_i + h)$  is the variable value at distance  $h$  from  $x_i$ . In this study,  $\gamma(h)$  estimates the variability of reflectance,  $Z$ , as a function of spatial separation.

The semivariogram has proven useful in remote sensing because it enables researchers to relate some of the descriptors of the semivariogram to the spatial characteristics of the scene. For example, the range, which defines the distance at which pixels are not spatially related, provides a measure of the size of the elements in the scene and has been suggested as a useful indicator in selecting the optimal spatial resolution for discriminating the features embedded in the image semivariogram (Curran, 1988). Woodcock *et al.* (1988) also found that the density of coverage of objects in the scene affects the height of the semivariogram. Atkinson and Danson (1988) used semivariograms to measure spatial dependence in coniferous and oak plantations. They found the range of the semivariogram was related to stand age and species, and were able to determine the optimal spatial resolutions for even-aged stands. Cohen *et al.* (1990) found the ranges for 1m spatial resolution data were related to the mean tree canopy sizes of the stands.

## STUDY AREA

The study area is located within the Central Plateau section of the Boreal Forest Region (Rowe, 1972) approximately 100 km north of Thunder Bay, Ontario, Canada. A diverse mosaic of forest stand types with various soil and landform conditions resulting from glacial activities are characteristic of the area. This area represents a low- to moderate-relief boreal environment with a rolling topography and terrain controlled by bedrock. Trembling aspen (*Populus tremuloides*) and black spruce (*Picea mariana*) are dominant with jack pine (*Pinus banksiana*), white spruce (*Picea glauca*), balsam fir (*Abies balsamea*), white birch (*Betula papyrifera*), white cedar (*Thuja occidentalis*) and tamarack (*Larix laricina*) occurring in various mixtures. Forest-stand overstories are monospecific or mixed and understories range from shrub- and/or herb-rich to poor.

The FEC that is applied in the study area is the Northwestern Ontario (NWO) FEC (Sims *et al.*, 1989). The framework upon which FECs are based incorporates those components of forest site which contribute to forest development (i.e., canopy and understory vegetation, soils, landform, climatic regime, and regional physiography) (Sims and Uhlig, 1992). FECs were developed for stand-level (i.e., <1:10,000) application to provide information regarding vegetation, soil and site conditions. The basic units of FECs are 'Vegetation Types' (V-Types) and 'Soil Types' (Sims *et al.*, 1989). Since FECs are hierarchical, field-level units can be aggregated to create lower-resolution ecosystem units (Hills and Pierpoint, 1960).

## METHODS

### Remote Sensing Data Acquisition

CASI data were acquired over the study site on July 30, 1993. These data were collected at approximately 10:30 a.m. along a flight line oriented parallel to the solar azimuth in order to minimize the effects of bidirectional reflectance (BDRF). A nine-band multispectral dataset was collected at two different altitudes with spatial resolutions of 1.39m x 5.36m (figures 1 and 2); and 0.73m x 5.36m (figures 3 and 4). The CASI data were converted to radiance and then to reflectance using procedures outlined in Babey and Soffer (1992) and Shepherd (1994).

## Semivariance Analysis of Multi-Spatial Resolution Remote Sensing Data

In this study, experimental semivariograms are used to model the underlying semivariogram for the CASI reflectance data. A single visible band and a near-infrared band were used for calculating semivariance and deriving semivariograms from low- and medium-altitude CASI data. Homogeneous stands (landscape units) of sufficient size were identified on aerial photographs and located on the low- and medium-altitude CASI images. To examine the spatial variability of forest ecosystems at high spatial resolutions, transects of 100 pixels in the cross-track direction were extracted from the CASI data for selected forest stands representing individual V-Types or complexes of V-Types (table 1). These stands are arranged in the table and in the corresponding figures along a gradient of pure hardwood to pure conifer. These visible and near-infrared data were subjected to semivariogram analysis<sup>1</sup> to estimate range and sill values using a spherical model. A series of transects were analyzed for each forest stand, from which the mean ranges and sills were calculated to characterize the stand / landscape unit.

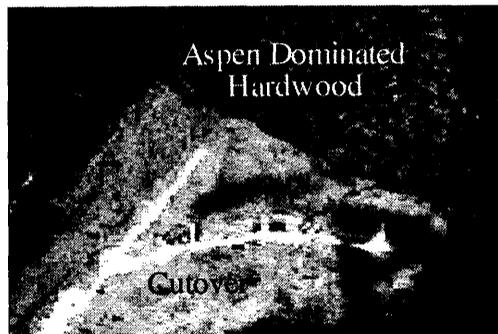


Figure 1.—Medium-altitude data

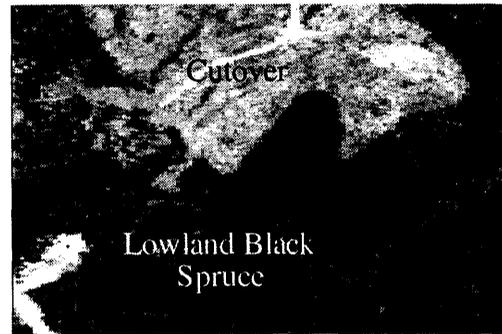


Figure 2.—Medium-altitude data

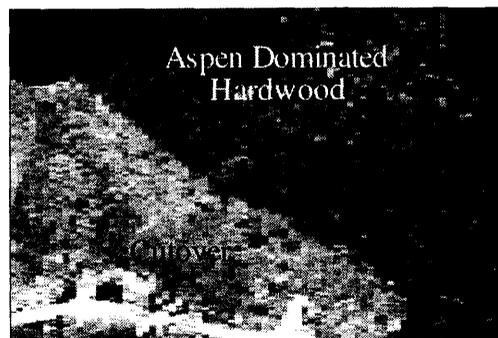


Figure 3.—Low-altitude data

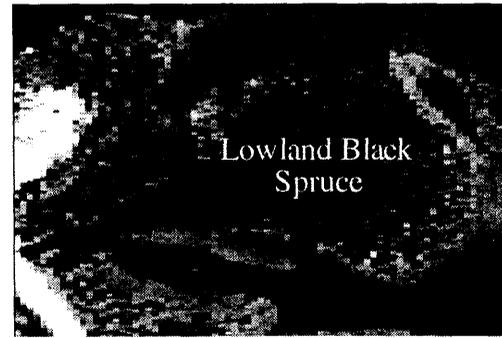


Figure 4.—Low-altitude data

<sup>1</sup> Geostatistical software: VARIOWIN (V.2.1) developed by Y. Pannatier at the Institute of Mineralogy of the University of Lausanne SWITZERLAND (PANNATIER, Y., 1994. MS-WINDOWS for Exploratory Variography and Variogram Modeling in 2D. In Proceedings, *Statistics of Spatial Processes: Theory and Applications*. Capasso, V., Girone G. and Posa, D. Eds., Bari, Italy, Sep. 27-30, 1993 pp. 165-170).

**Table 1.—Forest ecosystem groupings for semivariogram analysis**

<b>Forest Ecosystem Grouping</b>	<b>V-Types (defined by Sims <i>et al.</i>, 1989)</b>	
Aspen	V5	Aspen Hardwood
	V6	Trembling Aspen (W. Birch) - Balsam Fir / Mountain Maple
	V8	Trembling Aspen (W. Birch) / Mountain Maple
Aspen Complex	V9	Balsam Fir - White Spruce Conifer and Mixedwood
	V10	Trembling Aspen - Black Spruce - Jack Pine / Low Shrub
	V11	Trembling Aspen - Conifer / Blueberry / Feathermoss
Aspen Mixedwood	V8	Trembling Aspen (White Birch) / Mountain Maple
	V11	Trembling Aspen - Conifer / Blueberry / Feathermoss
	V19	Black Spruce Mixedwood / Herb Rich
Conifer Mixedwood	V8	Trembling Aspen (White Birch) / Mountain Maple
	V14	Balsam Fir Mixedwood
	V15	White Spruce Mixedwood
	V16	Balsam Fir - White Spruce Mixedwood / Feathermoss
	V19	Black Spruce Mixedwood / Herb Rich
Eastern White Cedar	V22	Eastern White Cedar
Upland Black Spruce	V33	Black Spruce / Feathermoss
Black Spruce Mix	V33	Black Spruce / Feathermoss
	V34	Black Spruce / Labrador Tea / Feathermoss (Sphagnum)
	V35	Black Spruce / Speckled Alder / Sphagnum
	V36	Black Spruce / Bunchberry / Sphagnum (Feathermoss)
Lowland Black Spruce	V35	Black Spruce / Speckled Alder / Sphagnum
	V36	Black Spruce / Bunchberry / Sphagnum (Feathermoss)
	V37	Black Spruce / Ericaceous Shrub / Sphagnum
Wetland Black Spruce	V37	Black Spruce / Ericaceous Shrub / Sphagnum
	V38	Black Spruce / Leatherleaf / Sphagnum

## RESULTS

At the low altitude, and for the stands sampled, the mean ranges derived from the semivariograms for the visible band indicate that trembling aspen stands have greater ranges ( $\cong 8$  pixels / 5.8m) than upland black spruce stands ( $\cong 6$  pixels / 4.4m) and lowland black spruce stands ( $\cong 4$  pixels / 2.9m) (figure 5). In general, the range values for the visible band increase with stand complexity. A similar trend is observed with respect to the range values for the near-infrared data whereby hardwood and conifer mixedwood stands have greater ranges than lowland black spruce. For example, range values are approximately 10 pixels / 7.3m for aspen complexes; 11 pixels / 8.0m for conifer mixedwood; 9 pixels / 6.6m for upland black spruce; and 6 pixels / 4.4m for lowland black spruce. The ranges derived for the near-infrared data are generally greater than those derived for the visible data for each ecosystem class (e.g., 6 pixels / 4.4m (infrared) versus 4 pixels / 2.9m (visible) for lowland black spruce). The sills for the trembling aspen and conifer mixedwood stands are much greater than those of the lowland black spruce stands, particularly for the near-infrared data. For example, the mean near-infrared semivariance (sill) is approximately 1700 for aspen stands; 2200 for conifer mixedwood stands; and 350 for lowland black spruce stands (figure 6). Although the trend is similar in the visible band, the differences are not as extreme.

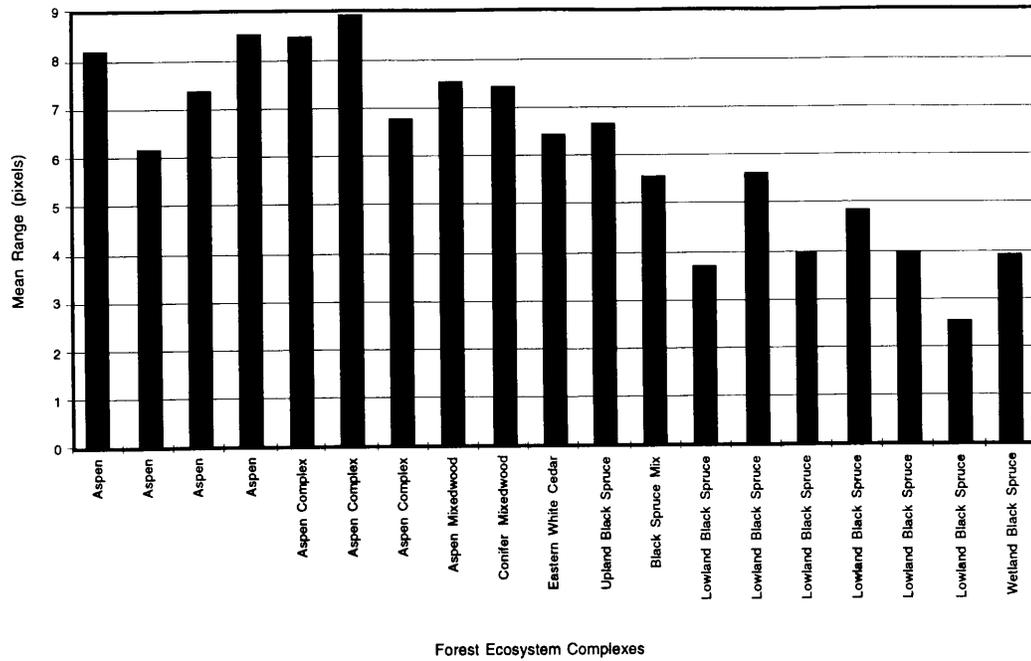


Figure 5.—Mean range values (580.06nm - 600.94nm) for forest ecosystem complexes in the Rinker Lake study area (Altitude = 600m)

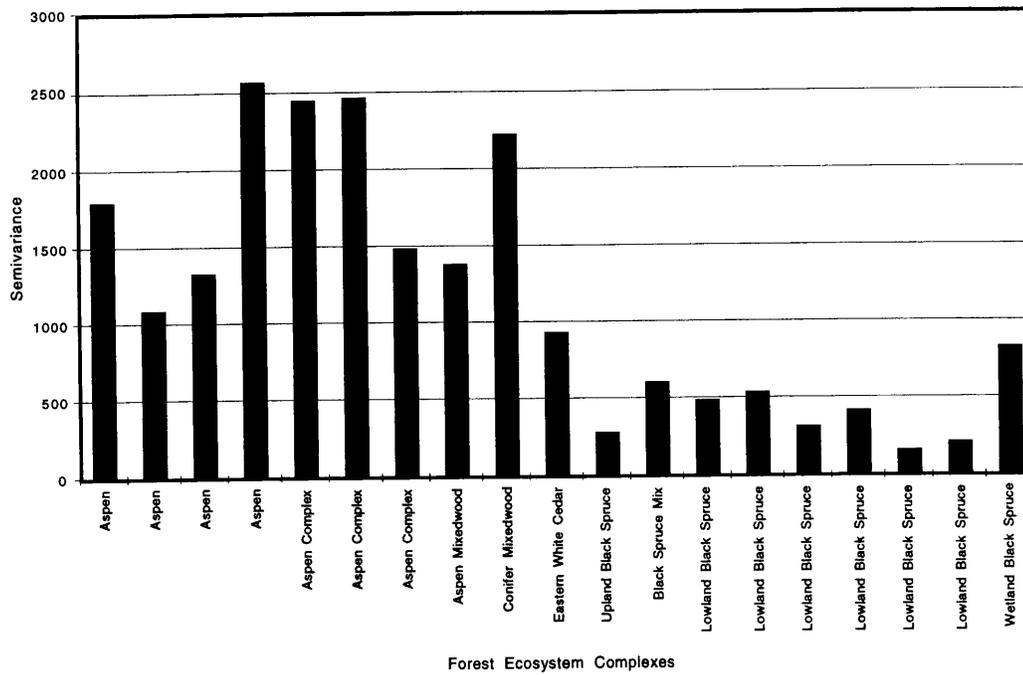


Figure 6.—Semivariance (743.64nm - 750.26nm) for forest ecosystem complexes in the Rinker Lake study area (Altitude = 600m)

For the medium-altitude data, the differences in mean ranges between the aspen stands and the lowland black spruce stands for the near-infrared data are not as pronounced. However, the ranges for the near-infrared data are greater than those for the visible data for all stands, particularly lowland black spruce (i.e., 7 pixels / 9.7m versus 3.5 pixels / 4.9m). Ranges derived from the medium-altitude data are greater than those measured for the low-altitude data. For example, the mean ranges for the trembling aspen stands are approximately 7.5 pixels / 5.5m for the low-altitude data while for the medium-altitude data they are approximately 6.5 pixels / 9.0m. Finally, the sills are typically lower for the medium-altitude data in comparison to the low-altitude data.

## CONCLUSIONS

For the low-altitude CASI data, the hardwood and mixed conifer stands have greater range values than pure conifer, particularly lowland conifer where canopy and understory are generally simple and homogeneous. At this resolution, individual trees may be the dominant feature affecting the semivariogram, particularly for the black spruce and trembling aspen stands since the crown diameters of these species are approximately equal to the ranges of influence. However, since these pixels are not square and individual trees are not observable, it is more likely that the ranges of influence are determined by an integration of a uniform number of tree crowns with understory and ground cover components.

The contrasting ranges of the near-infrared and visible data indicate that different features/phenomena are being measured in the near infrared as opposed to the visible, or at least the proportional contributions of features/phenomena to reflectance differ between these data. It is likely that the near-infrared data contain more ground-cover information since near-infrared energy has greater potential for penetration through a forest canopy. The differences in the range values observed at each spatial resolution for the same ecosystems reveals that information content also differs at the two altitudes. This observation would indicate that different features/phenomena or aggregations of features/phenomena (e.g., possibly 2-3 tree crowns with associated understories) are being measured at the medium resolution.

The observed differences in the magnitude of semivariance at both resolutions indicates that the trembling aspen stands (pure and mixed) and conifer mixedwood stands have increased layering with a greater volume of total cover than the black spruce stands. The complexity of the stand is more prevalent with the near-infrared data, since near-infrared energy has a greater capability of interacting with understory components. The occurrence of lower sills for the medium-altitude data indicates increased regularization as a result of decreased variability of forest stand reflectance at this resolution.

Based on these observations, it is anticipated that the potential for identifying forest ecosystems at a variety of levels will be improved by applying spatial analysis techniques (e.g., texture processing, spatial interpolation) to multi-spatial resolution remote sensing data. The identification of a single optimal spatial resolution for forest ecosystem discrimination may not be possible. Instead, it may be more suitable to analyze multi-spatial resolution data with appropriate processing techniques to extract various scales of information.

## ACKNOWLEDGMENTS

Funding for this project has been provided through the Northern Ontario Development Agreement, Northern Forestry Program. Additional support has been provided by a Center of Excellence Grant from the Province of Ontario to the Institute for Space and Terrestrial Science.

## REFERENCES

- Atkinson, P. and Danson, F. 1988. Spatial resolution for remote sensing of forest plantations, IGARSS'88 Symposium, Edinburgh, Scotland, ESA Publications Division, pp. 221-223.
- Babey, S. and Soffer, R. 1992. Radiometric calibration of the Compact Airborne Spectrographic Imager (CASI), *Canadian Journal of Remote Sensing*, 18(4):233-242.
- Curran, P.J. 1988. The semivariogram in remote sensing: an introduction, *Remote Sensing of Environment*, 24: 493-507.
- Hills, G. and Pierpoint, G. 1960. Forest site evaluation in Ontario, Research Report No. 42, Ontario Department of Lands and Forests, Toronto, Ontario, Canada, 63 pp.
- Jupp, D.L.B., Strahler, A.H. and Woodcock, C.E. 1988. Autocorrelation and regularization in digital images: I. basic theory, *IEEE Transactions on Geoscience and Remote Sensing*, 26(4):463-473.
- Matheron, G. 1963. Principles of geostatistics, *Economic Geology* 58:1246-1266.
- Rowe, J. 1972. Forest Regions of Canada, Publication No. 1300, Canadian Forestry Service, Department of the Environment, Ottawa, Ontario, Canada, 172 pp.
- Shepherd, P. 1994. Section 3.3, Airborne Remote Sensing Data, in Kalnins, V., Treitz P. and Howarth, P., Rinker Lake Data Report: 1993-1994, Earth-Observations Laboratory, Technical Report ISTS-EOL-TR94-002, pp. 19-48.
- Sims, R. and Uhlig, P. 1992. The current status of forest site classification in Ontario, *The Forestry Chronicle*, 68(1):64-77.
- Sims, R., Towill, W., Baldwin, K. and Wickware, G. 1989. Field Guide to Forest Ecosystem Classification for Northwestern Ontario, Ontario Ministry of Natural Resources, Toronto, Ontario, Canada, 191 pp.
- Woodcock, C.E., Strahler, A.H. and Jupp, D.L.B. 1988. The use of variograms in remote sensing: I. scene models and simulated images, *Remote Sensing of Environment*, 25:323-348.

## BIOGRAPHICAL SKETCH

Paul Treitz is a Lecturer in the Department of Geography (Faculty of Arts) and Faculty of Environmental Studies at York University, North York, Ontario, Canada. He is also a Ph.D. student in the Department of Geography at the University of Waterloo, Waterloo, Ontario, Canada.

Philip Howarth is a Professor of Geography at the University of Waterloo, Waterloo, Ontario, Canada. He is also a Principal Investigator with the Earth-Observations Laboratory of the Institute for Space and Terrestrial Science, a center of excellence established by the Province of Ontario for remote sensing and space related research.

# Application of Multiscale Modeling for Characterizing Variability in Hydrologic Processes

Praveen Kumar<sup>1</sup>

---

**Abstract.**—We describe a multiscale modeling framework applicable for a wide range of hydrologic processes. The foundation for this work has been laid by *Basseville et al.* [1992] and *Chou et al.* [1994]. Their development is based on treating the scale parameter akin to time, such that description at a particular scale captures the features of the process up to that scale that are relevant for the prediction of finer scale features, as in the notion of state-space approaches. We show the applicability of this model for modeling and estimation of a broad range of multiscale phenomena in hydrology. The methodology is illustrated through an example using a soil-moisture field obtained using remote sensing observations.

---

## INTRODUCTION

In varied guises throughout hydrologic science we encounter questions concerning the quantitative relationship between processes occurring at disparate spatial and temporal scales. These questions are mostly triggered by our need to understand, model and predict hydrologic processes over a broad range of spatial and temporal scales. Questions about multiscale manifestation of processes are at the heart of contemporary hydrologic research and answers to these questions will significantly advance our efforts in understanding the dynamics of the large scale hydrologic cycle.

Over the last several years there has been significant effort devoted to organizing field experiments in combination with remote sensing measurements, such as FIFE (*Sellers et al.*, 1992) and SIR-C [*Stofan et al.*, 1995], to improve our understanding of surface hydrologic and ecological processes and their impact on the atmosphere. In order to improve our understanding of a process at a particular scale we need to make measurements at that scale. Often it is difficult to decide the scale of significance and the strategy adopted is to make measurements at several scales with fine scale measurements of limited coverage embedded within coarse scale measurements of larger coverage (see Figure 1). Further analysis is performed to integrate the information across scales, to understand what area-averaged fluxes represent in conjunction with

---

<sup>1</sup>Department of Civil Engineering, University of Illinois, Urbana, Illinois 61801

point fluxes, and what physical controls on area-averaged fluxes need to be considered by a prognostic model [Smith *et al.*, 1992].

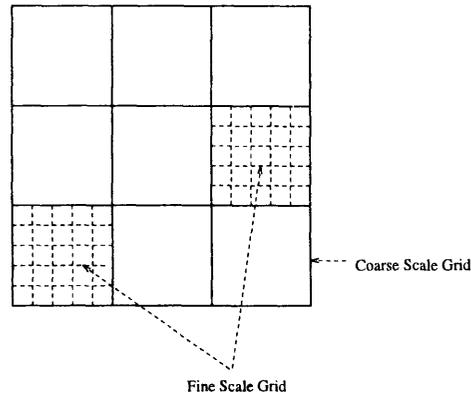


Figure 1: Schematic showing an idealized embedding of fine scale measurement grid of limited coverage within a coarse scale grid of larger coverage for a two level scheme.

Different techniques exist for analyzing observations obtained at different resolutions and are well explained in Daley [1991]. These range from simple function fitting and stochastic interpolation (univariate and multivariate) schemes to four dimensional data assimilation (FDDA). As indicated by Daley [1991], the essence of a good scheme is to account for the following factors simultaneously:

1. Observations have errors that can be temporally and spatially correlated with each other.
2. The process state variables are related to each other by the governing laws and are not independent.
3. There can be large data voids where no direct observations exist.

An ideal algorithm should filter observational noise, variance associated with scales too small to be properly resolved by the sampling scheme, and fluctuations that do not satisfy the governing laws.

Although multivariate statistical analysis or FDDA address these issues, *they do not include the crucial element of scale as an explicit parameter in the formulation.* This causes severe limitations in our ability to interpret the behavior of the phenomenon at different scales and the influence of the variability of one process at a small scale on the variability of a related process at a larger scale, except as a gross aggregate. For examples, Several studies (Mahrt [1987], Wetzel and Chang [1988], Avissar and Pielke [1989]) have indicated that grid

scale flux averages and sub-grid flux variance are both critical in characterizing the response of the atmosphere to surface fluxes. In addition no multiscale modeling scheme is immediately available from these formulations. In order to address these issues and more effectively utilize the observations obtained from the hierarchical sampling scheme (see Figure 1) to formulate conjectures and develop models we need an analysis-modeling-assimilation technique that is inherently *hierarchical in scale*.

Wavelet transforms have been significantly exploited in geophysics for the analysis of processes to infer their multiscale variability. They, however, are not easily amenable as a modeling tool. Motivated by the scale to scale hierarchical decomposition of wavelet transforms, significant advances in multiscale modeling have been made by *Basseville et al.* [1992] and *Chou et al.* [1994]. The key to their development is the treatment of the scale parameter akin to time, such that description at a particular scale captures the features of the process up to that scale that are relevant for the prediction of finer scale features, as in the notion of state-space approaches. These developments provide a natural setting not only for analysis and modeling of multiscale phenomena but also for multiscale data assimilation.

This paper attempts to provide more visibility to this multiscale modeling tool in the geophysics community through illustration using a simple example. In the next section, we briefly review the modeling framework and then provide an application to modeling soil moisture fields and estimating missing values.

### MULTISCALE STATE-SPACE MODEL

The following description briefly summarizes the basics of the ideas developed by *Basseville et al.* [1992] and *Chou et al.* [1992]. Consider the problem of disaggregating a random field (in two dimensions) from coarse to fine resolution. At the coarsest resolution the field will be represented by a single value (see Figure 2). At the next resolution there will be four values and in general, at the  $m^{\text{th}}$  resolution we obtain  $4^m$  values. The values of the random field can be described on the index set  $(m, i, j)$  where  $m$  represents the resolution and  $(i, j)$  the location index. The scale to scale decomposition can be schematically depicted as a tree structure (quadtree for two-dimensional processes) as shown in Figure 3. To describe the model let us use an abstract index  $\lambda$  to specify the nodes on the tree and let  $\gamma\lambda$  specify the parent node of  $\lambda$  (see Figure 3). Then the multiscale stochastic process can be represented as

$$X(\lambda) = A(\lambda)X(\gamma\lambda) + B(\lambda)W(\lambda). \quad (1)$$

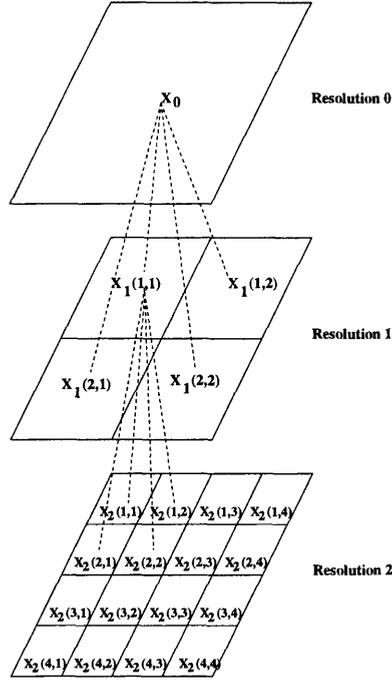


Figure 2: The structure of a multiscale random field is shown. The values at various grid locations  $(i, j)$  are given as  $x_m(i, j)$  where  $m$  is the resolution index. At the coarsest resolution ( $m = 0$ ), the field is represented by a single value or state vector, and generally at the  $m^{\text{th}}$  resolution there are  $4^m$  state vectors (scale and resolution are inversely related).

$A(\lambda)$  and  $B(\lambda)$  are matrices of appropriate sizes. We can assume that  $X_0 \equiv X(0)$  and  $W$  are normally distributed with known covariance matrices, i.e.,

$$X_0 \sim N(0, P_0) \quad P_0 = E[X(0)X^T(0)] \quad (2)$$

$$W(\lambda) \sim N(0, I) \quad (3)$$

where  $I$  is the identity matrix. Interpreting the states at a given level of the tree as a representation of the process at one scale, we see that equation (1) describes the evolution of the process from coarse to fine scale. *The term  $A(\lambda)X(\gamma\lambda)$  represents the interpolation or prediction down to the next finer level and  $B(\lambda)W(\lambda)$  represents new information added as the process evolves from one scale to the next.* The covariance  $P_\lambda \equiv E[X(\lambda)X^T(\lambda)]$  of the state at node  $\lambda$  evolves as

$$P_\lambda = A(\lambda)P_{\gamma\lambda}A^T(\lambda) + B(\lambda)B^T(\lambda). \quad (4)$$

Notice that the form of equations (1) and (4) do not change whether we deal with one-dimensional or multi-dimensional processes.

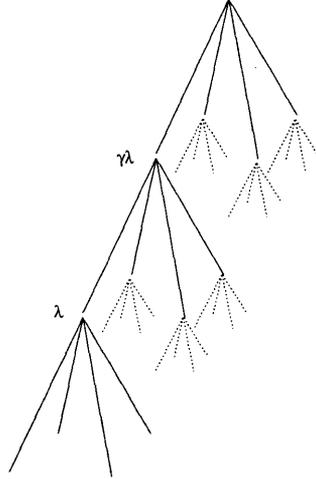


Figure 3: Multiscale decomposition. The abstract index  $\lambda$  refers to a node in the tree and  $\gamma\lambda$  refers to the parent node.

A particular form of model (1) is already popular as disaggregation model in hydrologic time series analysis [see *Valencia and Schaake, 1973*]. Equation (1), however, provides a more general modeling framework. For instance, it is possible to choose  $A(\lambda)$  and  $B(\lambda)$  such that (1) leads to a fractal model. For example, the model

$$X(\lambda) = X(\gamma\lambda) + B4^{-\beta \log_2 \lambda/2} W(\lambda) \quad \text{where } B \text{ is a constant} \quad (5)$$

leads to fractal processes with spectrum  $1/f^\beta$  [*Chou et al., 1992*] ( $\beta = 2$  corresponds to Brownian motion).

Multiscale stochastic models can be used to make estimates of processes from noisy measurements and for assimilation of data obtained at different resolutions. If we are given noisy measurements  $Y(\lambda)$  of the process  $X(\lambda)$  we can develop an estimation problem using the combination of equation (1) and the following equation

$$Y(\lambda) = C(\lambda)X(\lambda) + V(\lambda) \quad (6)$$

where  $V(\lambda) \sim N(0, R(\lambda))$ . The Matrix  $C(\lambda)$  can specify, in a very general way, measurements taken at different spatial locations and at different scales.  $R(\lambda)$  specifies the covariance of the measurement errors  $V(\lambda)$ . Equations (1) and (6) can be solved jointly to obtain estimates of  $X(\lambda)$ . Notice that this is a very attractive technique as *it enables us to combine estimation and filtering while exploiting hierarchical measurements at different scales*. Also, note that coarse scale observations provide the “context” or the background field for the next finer scale, i.e., scale plays a role akin to time in existing data assimilation schemes.

## APPLICATION TO A SOIL MOISTURE RANDOM FIELD

We describe the applicability of the multiscale model for modeling a soil moisture field and estimating missing values using a fractal prior model. The data, collected during the Washita' 92 Experiment [*Jackson et al.* 1993], were obtained using ESTAR – Electronically Steered Thinned Array Radiometer – images and converted to volumetric soil moisture content. These data sets, which correspond to several days of daily measurements of soil moisture in the top 5 cm layer, show scaling or fractal characteristics [*Rodriguez-Iturbe et al.*, 1995]. We study the field obtained on the second day of measurement (June 11, 1992) by extracting a sub-field with grid dimension  $64 \times 64$  (Figure 4), from the original of  $228 \times 93$ . As is evident from the Figure there are several missing values, indicated by dark patches, either due to the presence of township or roads, or measurement errors. For this study we do not distinguish between these and treat all as missing values due to measurement errors.

We use the results of analysis presented by [*Rodriguez-Iturbe et al.*, 1995] to estimate the parameter  $\beta = 0.79$  in equation 5. The multiscale model we use is give by the equations:

$$X(\lambda) = X(\gamma\lambda) + 57 \times 4^{-0.79 \log_2 \lambda / 2} W(\lambda) \quad (7)$$

$$Y(\lambda) = \begin{cases} X(\lambda) + V(\lambda) & \text{if } 2^\lambda = 64 \\ 0 & \text{otherwise,} \end{cases} \quad (8)$$

where  $V(\lambda = 6) \sim N(0, \sigma^2 = 25)$  gives the noise distribution at the finest scale. The algorithm first estimates the coarser scale values from the given finer scale values, i.e., going upward in the tree in Figure 3, and after reaching the root node proceeds downward to predict the estimates at all grid points and all scales, thereby providing estimates of the missing values. The results of the multiscale estimation are given in Figure 4. As is evident, the algorithm does an extremely good job of estimating the missing values. The estimation error are within the range of the data values.

The above is a simple example of the applicability of the multiscale model, although the model is vary general. In particular the state at any node can be described as a vector. For example, one may represent the state at any node by a vector whose elements are the components of the surface energy balance equation. The model can then be used to estimate the energy balance at various scales by combining point and remote sensing estimates. These issues are currently under investigation.

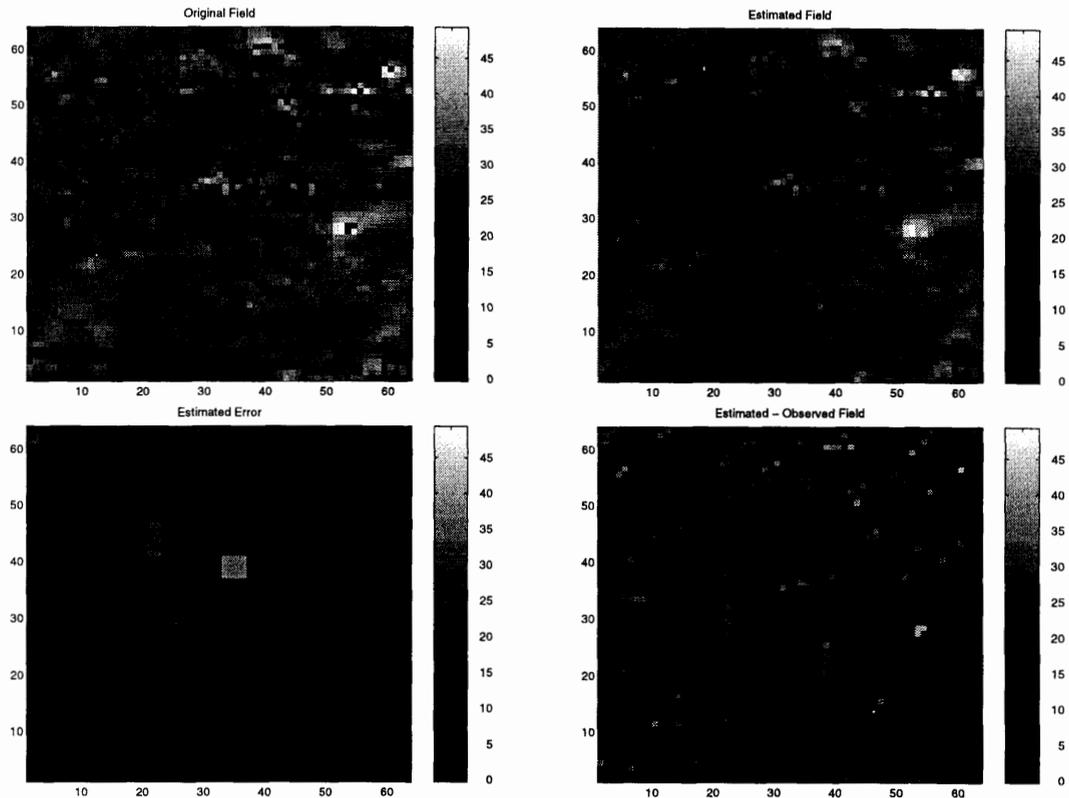


Figure 4: This figure shows the original volumetric soil moisture field (top-left) and the estimated field using a fractal prior model. The lower row shows the estimation error and the difference field.

## REFERENCES

1. Avissar, R. and R. Pielke, A parameterization of Heterogeneous Land-Surface for Atmospheric Numerical Models and its Impact on Regional Meteorology, *Mon. Weather Rev.*, 117, 2113-2136, 1989.
2. Basseville, M., A. Benveniste, and A. S. Willsky, Multiscale Autoregressive Processes, Parts 1 and 2, *IEEE Trans. on Signal Processing*, 40, 1915-1954, 1992.
3. Chou, K. C., A. S. Willsky, A. Benveniste, Multiscale Recursive Estimation, Data Fusion, and Regularization, *IEEE Trans. on Auto. Control*, 39(3), 1994.
4. Daley, R., *Atmospheric Data Analysis*, Cambridge University Press, 1991.
5. Mahrt, L., Grid-Averaged Surface Fluxes, *Mon. Weather Rev.*, 115, 1550-1560, 1987.

6. Rodriguez-Iturbe, I., G. K. Vogel, R. Rigon, D. Entekhabi, F. Castelli, A. Rinaldo, On the Spatial Organization of Soil Moisture Fields, *Geophysical Res. Letters*, 22(20), 2757-2760, 1995.
7. Sellers, P. J., F. G. Hall, G. Asrar, D. E. Strelbel, and R. E. Murphy, An Overview of the First International Satellite Land Surface Climatology Project (ISLSCP) Field Experiments (FIFE), *Jnl. of Geophysical Res.*, 97(D17), 18345-18371, Nov., 1992.
8. Smith, E. A. et al., Area-Averaged Surface Fluxes and Their Time-Space Variability Over the FIFE Experimental Domain, *Jnl. of Geophysical Research*, 97(D17), 18599-18622, 1992.
9. Stofan, E. R., et al., Overview of Results of Spaceborne Imaging Radar-C, X Band Synthetic Aperture Radar (SIR-C/X-SAR), *IEEE Trans. on Geoscience and Remote Sensing*, 33(4), 817-828, 1995.
10. Valencia, D., and J. C. Schaake, Jr., Disaggregation Processes in Stochastic Hydrology, *Water Resources Res.*, 9(3), 580-585, 1973.
11. Wetzal, P. J., and J. T. Chang, Evapotranspiration from Nonuniform Surfaces: A First Approach for Short-Term Numerical Weather Prediction, *Mon. Weather Rev.*, 116, 600-621, 1988.

### **BIOGRAPHICAL SKETCH**

Praveen Kumar is an Assistant Professor in the Department of Civil Engineering at the University of Illinois in Urbana-Champaign, Illinois. He obtained his PhD from the University of Minnesota in 1993. His research interests include analysis and modeling of multiscale variability in surface hydrologic processes, large scale hydrologic processes, and hydrometeorology and hydroclimatology.

# Flow Based Scale-up of Heterogeneous Porous Media using Homogenization and Wavelet Representation

Joe Koebbe<sup>1</sup> and Ryan Thomas<sup>2</sup>

---

**Abstract** - Scaling up microscopic heterogeneities to a macroscopic level useful in reservoir engineering applications is a difficult problem. Many of the small scale heterogeneous structures such as shale barriers that occur deltaic rock formations dictate flow on the larger scale. Most averaging methods are not able to maintain the information contained in these types of structures at the macroscopic scale. Another important example is that of a fractured reservoir where depending on age, fractures can act as either conduits or barriers to flow through the rock. The goal of the research presented is to use (1) flow based averaging techniques to construct a more rigorous way of averaging the rock properties and (2) use wavelet representation in the averaging method to preserve the effect of any microscopic structures that strongly influence the flow. Results from the application of this methodology to the rock formations at the Ferron Sandstone site in southern Utah will be presented. Transect data obtained by the Utah Geological Survey in joint work with Mobil Oil, British Petroleum, and a number of other research groups will be used as a basis of the numerical studies.

---

## SCALING PROBLEMS IN POROUS MEDIA

Computer modeling of fluid flow in porous media is used to predict the performance of reservoirs in groundwater flow and petroleum engineering applications. In many cases the porous medium is highly heterogeneous. The heterogeneity occurs at all scales in the problem and very small scale features of a reservoir can strongly influence flow of fluids in the reservoir. In most applications, computer models are needed on a scale that is large relative to many important geological features. Exactly resolving the important small

---

<sup>1</sup>Associate Professor, Department of Math & Stat, Utah State University, Logan, UT

<sup>2</sup>Undergraduate Student, Department of Math & Stat, Utah State University, Logan, UT

scale features is impossible and normally some method of averaging is used in the hope that the bulk or averaged properties will still contain the important information from the small scale heterogeneities.

For example, fluvial deltaic formations, which comprise a large share of the known hydrocarbon bearing reservoir formations in the world, are created by patterns of depositional events that leave possibly large deposits of coarse grain material such as sand, separated by small amounts of finer grain materials, such as clays and silt. The resulting rock formation from these types of depositional environments is layered with high permeability regions separated by thin, relatively low permeability layers. The bulk flow in these formations is strongly influenced by these very small scale features.

Fractures in a formation are also small scale features that can function as either conduits or barriers. Just after formation a fracture will function as a conduit allowing fluid to move from one region to another easily. Later as minerals deposit along the fracture surface the fracture becomes a barrier to flow. Each of these regimes is important in reservoir modeling. To understand how a formation has trapped the hydrocarbons it necessary to be able to model the early flow history through the fracture and to understand how to produce the hydrocarbon effectively after the fractures have shut off, the behavior of the fracture as a barrier to flow must be understood. The effects of these relatively small scale heterogeneities must be included in the computer model.

## AVERAGING PROCESSES AND HOMOGENIZATION

The description will focus on a simple flow model, Darcy's equation, for a porous medium. The simplest partial differential equation that models flow in a reservoir is

$$\nabla \cdot \mathbf{K}(\mathbf{x})\nabla h(x) = f(x) \quad (1)$$

The parameters in the problem are the rock permeability (or conductivity) tensor,  $\mathbf{K}$  defined by

$$\mathbf{K} = k_{xx}, \quad \mathbf{K} = \begin{bmatrix} k_{xx} & k_{xy} \\ k_{yx} & k_{yy} \end{bmatrix}, \quad \mathbf{K} = \begin{bmatrix} k_{xx} & k_{xy} & k_{xz} \\ k_{yx} & k_{yy} & k_{yz} \\ k_{zx} & k_{zy} & k_{zz} \end{bmatrix}$$

in one two and three dimensions, respectively, the pressure head (or hydraulic head)  $h$ , the spatial variable  $x$ , and a forcing function  $f$  which can be used to represent wells or confining reservoirs. For simplicity assume that the permeability varies on only two scales which will be termed the microscopic and macroscopic scales. In what follows the only parameter that will be scaled up will be the permeability,  $K$ . For convenience the field is assumed to be periodic. The ideas here have been extended to aperiodic permeability fields.

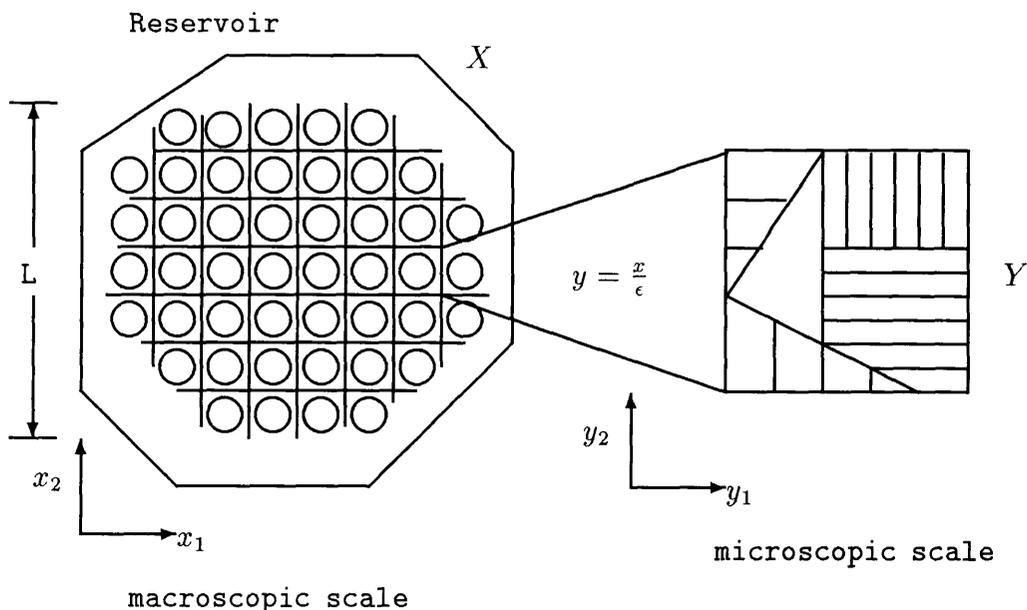


Figure 1: Illustration of the (a) macroscopic scale and (b) a microscopic view of a cell.

Figure (1) depicts the two length scales in two spatial dimensions. The pattern of the permeabilities from a distant or macroscopic view is periodic. As we zoom in on the cells we see that in each cell the structure of the permeability can be quite general. It is assumed that the cell is small compared to the overall size of the reservoir. Next define two length scales; the macroscopic scale represented by the variable  $x$  and the microscopic scale represented by the variable  $y$ . The macroscopic scale involves the pressure and velocity variables over the entire reservoir,  $X$ . The microscopic scale is defined by a single cell in Figure (1), denoted by  $Y$ . The relationship between  $x$  and  $y$  is  $y = x/\epsilon$ . Note that that on the microscopic scale as  $y$  varies from 0 to 1 the macroscopic variable  $x$  varies from 0 to  $\epsilon$ .

The goal is to determine a bulk permeability on the region  $X$  by scaling up the information in the small cells. There are some simple cases that can be discussed without complicated averaging. Consider a model of flow through a layered porous medium with well defined discontinuities between the different layers of rock. Assume the permeability in each type of rock is constant. If we assume that the fluid flow is either parallel or perpendicular to the direction of flow then the scale up methods that are appropriate are clear. When the flow is parallel to the discontinuities the arithmetic average should be used for the permeability and when the flow is perpendicular to the direction of flow the

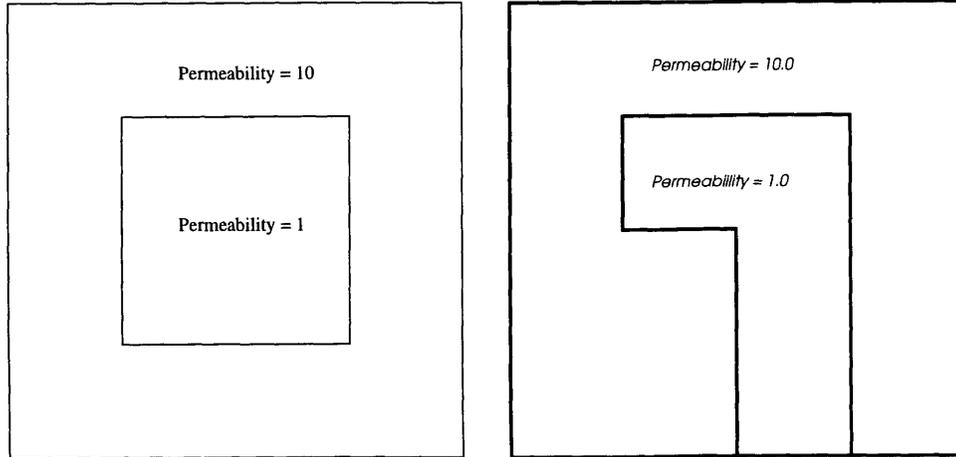


Figure 2: The symmetric and inverted-L patterns of permeabilities used for illustrating problems with functional averaging methods. The harmonic, arithmetic and geometric means return diagonal tensors.

harmonic average of the permeabilities should be used. With a little analysis it is easy to show that if the flow occurs at some other angle of incidence to the discontinuities the permeability should be between the harmonic and arithmetic average. The question is what is the appropriate effective value to be used based on the flow? In many cases ad hoc methods are used that combine harmonic and arithmetic averaging. If the field is log-normal the geometric average may be appropriate.

Each of these averaging methods is a functional averaging method. That is, no matter where the signal or data has come from the same averaging technique will be applied. How the data is used in the solution of the flow problem is not considered. Thus the standard methods of averaging are not conditioned on the underlying physical process. As an illustration of the problem consider the two permeability patterns shown in Figure (2). A symmetric and nonsymmetric pattern that looks like an inverted 'L' are shown. Each of these regions can be divided into 16 equally sized cells; 4 with permeability 1 and 12 with permeability 10. In two dimensions the permeability tensors in this case would be

$$\mathbf{K} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mathbf{I} \quad \text{or} \quad \mathbf{K} = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix} = 10 \mathbf{I}$$

where  $\mathbf{I}$  is the usual 2x2 identity matrix. Suppose that we impose a pressure gradient so that the fluid will flow from lower left to upper right in the region. We should expect the fluid to move around the obstacle in both case. the flow should be symmetric in the first case while asymmetric in the inverted-L case.

If we average the permeability tensor using any functional based method we will obtain a symmetric tensor which is a multiple of the identity. The tensors for the arithmetic, harmonic, and geometric averages are  $K_A = 7.75 \mathbf{I}$  and  $K_H \approx 3.077 \mathbf{I}$ , and  $K_G \approx 5.623 \mathbf{I}$ , respectively for both the symmetric and asymmetric patterns. You should note that in all of these cases the principle flow directions are still lined up along the coordinate axes. Thus using these averaging methods will produce the same tensor and the same bulk properties. We need an averaging method that is conditioned to the flow problem we are trying to solve.

The homogenization procedure explained below will generate tensors

$$\mathbf{K}^\# \approx \begin{bmatrix} 6.651 & 0 \\ 0 & 6.651 \end{bmatrix} \quad \text{and} \quad \mathbf{K}^\# \approx \begin{bmatrix} 5.574 & -0.348 \\ -0.348 & 6.949 \end{bmatrix}$$

for the symmetric and asymmetric patterns, respectively. For the symmetric pattern the principle flow directions are lined up along the coordinate axes as one would expect. In the asymmetric pattern the principle flow directions are at approximately 76 degrees from the horizontal axis.

In the next section a very brief description of the homogenization process will be given. The interested reader is referred to Bourgeat, 1984, for theoretical results and to Amaziane, et. al., 1990 for computational aspects of homogenization. In addition, a code exists for applying homogenization on rectangular grids Koebbe, 1996. The computer code can be downloaded via anonymous ftp from

`sunfs.math.usu.edu/pub/koebbe/homog.tar.Z`

and then uncompressed on a Unix operating system.

### The Perturbation Method

The averaging is done over one of the small cells as depicted in Figure (1). In the work below we will want to use the Darcy velocity in the computations which is defined by  $\mathbf{v} = -\mathbf{K}\nabla h$ . This quantity can be used to reduce the second order partial differential equation to a system of first order partial differential equations of the form

$$-\nabla \cdot \mathbf{v} = f \tag{2}$$

$$\mathbf{v} = -\mathbf{K}\nabla h. \tag{3}$$

The perturbation method will be applied to the system of partial differential equations defined by Equations (2) and (3).

The parameter  $\epsilon$  which is related to the size of the microscopic cell is small; that is  $0 < \epsilon \ll 1$ . The perturbation method then assumes that the

variables,  $h$  and  $\mathbf{v}$ , can be expanded in a series in the perturbation parameter  $\epsilon$ ;  $h = h_0 + \epsilon h_1 + \epsilon^2 h_2 + \dots$  and  $\mathbf{v} = \mathbf{v}_0 + \epsilon \mathbf{v}_1 + \epsilon^2 \mathbf{v}_2 + \dots$ . The spatial differentiation operator  $\nabla$  becomes  $\nabla = \nabla_x + \frac{1}{\epsilon} \nabla_y$  where the subscripts denote differentiation with respect to the specified variable. The differentiation is decomposed into two pieces; one at the macroscopic scale,  $\nabla_x$ , and one at the microscopic scale,  $\nabla_y$ . The microscopic derivative is scaled by  $\frac{1}{\epsilon}$  due to the relationship  $y = \frac{x}{\epsilon}$  mentioned above.

The multiple scale differentiation operator and the series expansions are then substituted into the first order system in Equations (2) and (3). The result is the system

$$-(\nabla_x + \frac{1}{\epsilon} \nabla_y)(\mathbf{v}_0 + \epsilon \mathbf{v}_1 + \dots) = f \quad (4)$$

$$\mathbf{v}_0 + \epsilon \mathbf{v}_1 + \dots = -\mathbf{K}^\epsilon (\nabla_x + \frac{1}{\epsilon} \nabla_y)(h_0 + \epsilon h_1 + \dots) \quad (5)$$

The perturbation method equates the coefficients multiplying the powers of the small parameter  $\epsilon$  on either side of Equations (4) and (5).

The equations for the powers of  $\epsilon^{-1}$  and  $\epsilon^0$  are:

$$-\nabla_y \cdot \mathbf{v}_0 = 0 \quad (6)$$

$$0 = \mathbf{K} \nabla_y h_0 \quad (7)$$

and

$$-(\nabla_x \cdot \mathbf{v}_0 + \nabla_y \cdot \mathbf{v}_1) = f \quad (8)$$

$$-\mathbf{v}_0 = \mathbf{K}(\nabla_x h_0 + \nabla_y h_1), \quad (9)$$

respectively. The coefficients that are multiplied by a power of  $\epsilon$  greater than zero are neglected. From this point on the perturbation is routine and the interested reader can check the references for the details of the analysis.

### Computation of the Homogenized Permeability

In the perturbation analysis it is necessary to assume the microscopic pressure variable  $h_1$  can be written as a combination of the partial derivatives of the macroscopic pressure variable  $h_0$  in the following way.

$$h_1(x, y) = \sum_{k=1}^3 w_k(y) \frac{\partial h_0}{\partial x_k}. \quad (10)$$

The multipliers  $w_k(y)$  allow the computation of a scaled combination of macroscopic derivatives and can be thought of as containing the first order fluctuations of the pressure on a microscopic cell,  $Y$ .

With this assumption the perturbation analysis results in an averaged first order system of equations

$$-\nabla_x \cdot \mathbf{v}^\# = f^\# \quad (11)$$

$$\mathbf{v}^\# = -\mathbf{K}^\# \nabla h_0 \quad (12)$$

These averaged or macroscopic equations are of the same form as the original equations defined for the original heterogeneous medium.

The effective permeability,  $\mathbf{K}^\#$ , returned by this perturbation analysis called the homogenized permeability is defined by

$$\mathbf{K}^\# = \langle \mathbf{K} \left( \mathbf{I} + \begin{bmatrix} \frac{\partial w_1}{\partial y_1} & \frac{\partial w_2}{\partial y_1} & \frac{\partial w_3}{\partial y_1} \\ \frac{\partial w_1}{\partial y_2} & \frac{\partial w_2}{\partial y_2} & \frac{\partial w_3}{\partial y_2} \\ \frac{\partial w_1}{\partial y_3} & \frac{\partial w_2}{\partial y_3} & \frac{\partial w_3}{\partial y_3} \end{bmatrix} \right) \rangle_Y \quad (13)$$

where the notation  $\langle \cdot \rangle$  indicates an integral average over the domain. The integral average is done for each of the nine components separately. The problem that we are left to solve is: Find the functions  $w_k(y)$  that are included in the form given in Equation (13).

If we substitute the form in Equation (13) into Equation (10) we generate a system of uncoupled elliptic partial differential equations of the form

$$\nabla_y \cdot \mathbf{K} \nabla w_k(y) = -\nabla \mathbf{K} \mathbf{e}_k \quad (14)$$

where  $k = 1, 2, 3$  and  $\mathbf{e}_k$  is the  $k^{th}$  unit vector in  $\mathfrak{R}^3$  (or  $\mathfrak{R}$  or  $\mathfrak{R}^2$  in the one and two dimensional cases, respectively). The solutions of these equations is done via some standard finite element method. For a piecewise constant permeability a piecewise continuous linear finite element will return exact results and also the continuity of flux condition is a natural boundary condition for the Galerkin method.

## SPECTRAL SOLUTIONS AND WAVELETS

To illustrate the connection of the homogenization process to wavelet representation consider the one dimensional restriction of Equation (14) to the unit interval  $[0, 1]$ . The homogenized permeability is computed using

$$\mathbf{K}^\# = \int_0^1 \mathbf{K} \left( 1 + \frac{dw}{dy} \right) dy \quad (15)$$

where the function  $w(y)$  must satisfy the ordinary differential equation

$$\frac{d}{dy} \mathbf{K} \frac{dw}{dy} = -\frac{d}{dy} \mathbf{K}$$

Physically,  $\mathbf{K}^\#$  should be the harmonic average and it is always the case that the homogenization process agrees with this value.

The problem can be viewed from the point of Sturm-Liouville theory if we include the boundary conditions. The conditions that are imposed by the physical problem are (1) continuity of the fluid flux, (2) the average returned must be the harmonic average, and (3) a 'wavelet' condition which can be stated as  $w(0)$ . Solving Equation (15) with the given conditions will produce a set of orthogonal eigenfunctions (trigonometric functions) that can be used as a basis for constructing characterizations of the permeability field from the point of view of Fourier analysis or wavelets. Details of this process will be presented in the talk and appear in another paper.

## REFERENCES

- Bourgeat, A., 1984 Homogenization method applied to the behavior of a naturally fissured reservoir, In K.I. Gross, editor, *Mathematical Methods in Energy Research*, pages 181–193. SIAM.
- Amaziane, B., Bourgeat, A., and Koebbe, J., 1990, Numerical simulation and homogenization of two-phase flow in heterogeneous porous media, Hornung, Dogan, and Knaber, editors, *Transport in Porous Media II*. Kluwer Academic Publishers.
- Koebbe, J.V., 1996, Homcode: A code for scaling up permeabilities using homogenization, (to appear as a Utah Geological Survey report).
- Bourgeat, A.P., and Hidani A., 1994, Effective model of two-phase flow in a porous medium of different rock types. *Publication de l'Equipe d'Analyse Numerique*, Lyon-St. Etienne
- Bourgeat, A.P., Kozlov, S.M., and Mikelic, A., April 1993, Effective equations of two-phase in random media, *Publication de l'Equipe d'Analyse Numerique*, Lyon-St. Etienne

## BIOGRAPHICAL SKETCH

Joe Koebbe is an Associate Professor in the Department of Mathematics and Statistics at Utah State University. He received a PhD in Mathematics from the University of Wyoming in 1988. Joe works in reservoir simulation with applications in groundwater flow and petroleum applications.

Ryan Thomas is an undergraduate student at Utah Stat University working towards a Bachelors degree in mathematics and Masters degree in electrical engineering. His area of interest is currently in signal processing.

# Nonhomogeneous Hidden Markov Models Allowing Stochastic Downscaling of Synoptic Atmospheric Patterns to Local Hydrologic Phenomena

Peter Guttorp<sup>1</sup> and James P. Hughes<sup>2</sup>

---

**Abstract.**— A model for multistation precipitation, conditional on synoptic atmospheric patterns, is presented. It postulates the existence of an unobserved weather state which serves as a link between the large-scale atmospheric measures and the small-scale spatially discontinuous precipitation field. The weather state process is assumed to be conditionally Markov, given the atmospheric data. The rainfall process is then conditionally specified given the weather state. Various parameterizations for the weather state process and the rainfall process are discussed and a likelihood based estimation procedure is described. As an example the model is fit to a twenty station network of rain gauge stations in western Washington. We discuss the NHMM as a method of relating synoptic atmospheric data, such as the output of a general circulation model, to rainfall and other local (sub-grid scale) hydrologic processes, and outline possible extensions to continental scales.

---

## INTRODUCTION

Stochastic downscaling of GCM data involves building a stochastic model to relate historical measurements of synoptic circulation to historical regional or local climate data (e.g. precipitation). Then, using GCM circulation data as input, the model can be used to produce simulations at the regional or local level. The strength of this procedure lies in its relative simplicity and computational ease compared to the difficult task of building a nested GCM. However, the procedure assumes that the relationships (between circulation and regional climate) identified in the historical data will be preserved under the alternative climate scenarios produced by the GCM.

Initial attempts at stochastic downscaling utilized “weather state” models (e.g. McCabe et al. (1989), Hay et al. (1991), Wilson et al.(1992), Bardossy and Plate (1992)). In these models, each day is explicitly classified into one of a few

---

<sup>1</sup> *Department of Statistics, Box 354322, University of Washington, Seattle, WA 98195-4322.*

<sup>2</sup> *Department of Biostatistics, University of Washington, Seattle, WA.*

discrete categories (weather states) based on the synoptic circulation pattern for that day. Then separate rainfall models are fit conditional on the weather state. Hughes et al. (1993) developed such a model using historical circulation and rainfall and then applied the results to data from a GCM  $2 \times \text{CO}_2$  run to describe rainfall patterns, streamflow and flood frequency under the altered climate.

A difficulty with the weather state models as defined above is that the weather states must be explicitly defined a priori. A poor choice of the weather states will lead to a poor model. As an alternative, Zucchini and Guttorp (1991) proposed a hidden Markov model (HMM) as a model for precipitation. The HMM presupposes the existence of a few discrete weather states but it assumes that these states cannot be observed directly. Instead, only rainfall is observed and the rainfall is assumed to be conditionally temporally independent given the weather state. In this model the weather states are effectively defined by the rainfall process. The states chosen are those that separate the rainfall process into different spatial or temporal patterns. Formally, let  $\mathbf{R}_t$  be the measurement (typically multivariate) of the regional or local process (e.g. rainfall occurrence) at time  $t$  and  $S_t$  be the weather state at time  $t$ . Then the hidden Markov model is defined by the assumptions

$$P(\mathbf{R}_t | S_1^T, \mathbf{R}_1^{t-1}) = P(\mathbf{R}_t | S_t) \quad (1)$$

$$P(S_t | S_1^{t-1}) = P(S_t | S_{t-1}) \quad (2)$$

where the notation  $S_1^T$  means all values of  $S_t$  from time 1 to  $T$ . The term hidden refers to the fact that the weather state process,  $S_t$ , is an unobserved quantity and is not defined a priori (although the weather state for each day may be identified after the model is fit). Nonetheless, Zucchini and Guttorp (1991) found that the circulation patterns associated with particular weather states were interpretable. In effect, the model acts as an automatic classifier of circulation patterns into weather states that are associated with particular precipitation patterns.

The major drawback of the HMM approach is that it fails to incorporate synoptic circulation information in the definition of the weather states. The weather states are defined solely on the basis of rainfall. Thus, the model cannot be used for downscaling. However, an extension to the HMM which does incorporate atmospheric information and can be used for downscaling was developed by Hughes and Guttorp (1994a). They termed their model a nonhomogeneous hidden Markov model (NHMM). Letting  $\mathbf{X}_t$  be the measurement (or summary) of the atmospheric data at time  $t$ , the NHMM is defined by the assumptions

$$P(\mathbf{R}_t | S_{1^t}, \mathbf{R}_1^{t-1}, \mathbf{X}_1^T) = P(\mathbf{R}_t | S_t) \quad (3)$$

$$P(S_t | S_1^{t-1}, \mathbf{X}_1^T) = P(S_t | S_{t-1}, \mathbf{X}_t) \quad (4)$$

The first assumption says that the local or regional process,  $\mathbf{R}_t$ , is conditionally temporally independent given the current weather state. In other words, the regional process is driven by the weather state process. The second assumption

states that the (unobserved) weather state process is Markov with transition probabilities depending on the current values of the atmospheric information. Various parameterizations are possible for  $P(S_t|S_{t-1}, \mathbf{X}_t)$  and  $P(\mathbf{R}_t|S_t)$  and maximum likelihood techniques are available to estimate the parameters. Hughes and Guttorp (1994a) show that both the explicit weather state models mentioned previously and the hidden Markov model are special cases of the more general NHMM.

Hughes and Guttorp (1994a) applied the NHMM to 24 years of winter data from a four station rain gauge network in Washington state using the following model for  $P(\mathbf{R}_t|S_t)$ :

$$P(\mathbf{R}_t = \mathbf{r}_t | S_t = s) = \prod_{i=1}^n p_{is}^{r_i^i} (1 - p_{is})^{1-r_i^i} \quad (5)$$

This model assumes that the probability of rain at each station depends only on the weather state and is independent of the other stations. For the network analyzed by Hughes and Guttorp (1994a) this assumption was reasonable since the rain stations were widely separated. Note that such an assumption does not imply that the rainfall occurrence processes at the stations are uncorrelated. Hughes and Guttorp (1994a) found strong (unconditional) correlations, on the order of 0.4 to 0.5, between the stations. This correlation is induced by the common weather state.

In Hughes and Guttorp (1994b), the NHMM was applied to 21 years of winter data in a densely packed network of 24 stations in the Puget Sound region of western Washington state. The assumption of spatial independence (even conditional on the weather state) was no longer tenable. Therefore, the following model was used for  $P(\mathbf{R}_t|S_t)$ :

$$P(\mathbf{R}_t = \mathbf{r}_t | S_t = s) \propto \exp\left(\sum_{i=1}^n \alpha_{is} r_i^i + \sum_{j < i} \beta_s r_i^i r_j^j / d_{ij}\right) \quad (6)$$

where  $d_{ij}$  is the distance between stations  $i$  and  $j$ . The parameters  $\beta_s$  measure the spatial dependence between stations. When  $\beta_s$  is positive, stations  $i$  and  $j$  are positively correlated (within weather states) and the correlation decreases inversely with distance. A negative value for  $\beta_s$  implies negative correlation between stations  $i$  and  $j$  (within weather states). When each  $\beta_s$  is 0, equation (6) reduces to the independence model (equation (5)) with  $\alpha_{is} = \log(p_{is}/(1 - p_{is}))$ . As before, maximum likelihood may be used for parameter estimation. However, for large numbers of stations the constant of proportionality in equation (6) is difficult to evaluate, and Hughes and Guttorp (1994b) describe an alternative estimation strategy for that situation.

When applied to the 24 station network mentioned above, and based on mean sea-level pressure and the north-south gradient of 500 mb geopotential height, the model identified five weather states. State 1 had a southwesterly flow over the Puget Sound region, both at the surface and aloft, resulting in high precipitation probabilities everywhere. State 2 had westerly surface flow, but southwest flow aloft. There was substantial rainfall in the north and south, but relatively

low probability of rain in the central part of the region. State 3 was characterized by southwesterly flow at sea level and westerly flow aloft, with fairly weak surface winds. The stations bordering Puget Sound were in a rain shadow, while more easterly stations had high precipitation probabilities. State 4 was characterized by a high pressure system over the region, with consequent low precipitation rates. Finally, state 5 was similar to state 2, but with a more westerly 500 mb flow. The rainfall pattern was similar to that in state 2.

Statistics of the historical rainfall process (i.e. probability of rain, spatial and temporal correlations, storm duration distribution) were generally well reproduced although stations at the periphery of the network or isolated stations tended to exhibit a predicted (simulated) duration distribution which was lighter-tailed than that which was actually observed.

### MODELING PRECIPITATION ON LARGER SCALES

The NHMM has not yet been applied on subcontinental or continental scales. Some modification of the model will be necessary in this case since, in the present configuration,  $S_t$  is a finite scalar quantity which is meant to describe the weather state at time  $t$  over a (meteorologically) homogeneous region. One possibility would be to identify  $K$  distinct precipitation/meteorological regions and then allow  $S_t$  to be a  $K$ -variate vector where the  $k$ -th element corresponds to the weather state in the  $k$ -th region. The model for  $P(S_t|S_{t-1}, X_t)$  therefore becomes a model for a multivariate Markov process. The simplest such model is

$$P(S_t|S_{t-1}, X_t) = \prod_i P(S_t^i|S_{t-1}^i, X_t) \quad (7)$$

which postulates that each regional process evolves independently, conditional on the observed atmospheric process,  $X_t$ . Presumably, one would also want to assume that the rain stations in each region depend only on that region's hidden state:

$$P(R_t|R_t^{(i)}|S_t^i) \quad (8)$$

where  $R_t^{(i)}$  denotes the set of rain stations in region  $i$ . Note that this assumption of spatial independence between regions does not preclude the use of a model for spatial dependence within each region.

Unfortunately, it seems unlikely that the situation will be as simple as the (conditional) regional independence model envisioned in (1). It is possible that some enhancement in skill can be gained by coupling weather states in each region with weather states in other regions. For instance, the movement of large scale storm fronts across regions would suggest that yesterday's weather state in one region could provide information about today's weather state in another region. On the other hand, if  $N = \prod_{k=1}^K \dim(S^k)$  is the total number of states in the  $K$ -variate process  $S_t$ , it is not computationally feasible to attempt to fit the entire  $N(N-1)$  transition matrix,  $P(S_t|S_{t-1}, X_t)$ . A reasonable compromise between these extremes might be to assume that today's state in region  $i$  depends only on

yesterday's state in the regions surrounding region  $i$ . That is,

$$P(S_t|S_{t-1}, X_t) = \prod_i P(S_t^i|S_{t-1}^{\mathcal{N}(i)}, X_t) \quad (9)$$

where  $\mathcal{N}(i)$  is read "the neighborhood of  $i$ ". Basic meteorologic principles and knowledge of the regions under consideration should be used to determine which regions constitute the neighborhood of region  $i$ .

Another approach to modeling continental scale precipitation would be to divide the atmospheric variables into region-specific measures (call these  $X_t^{R_i}$  for region  $i$ ) and (sub)continental measures (call these  $X_t^C$ ) and to postulate the existence of an unobserved continental scale weather state (call this  $C_t$ ). Under these assumptions, one might write

$$P(S_t|S_{t-1}, X_t) = \sum_{C_t} P(C_t|X_t^C) \prod_i P(S_t^i|S_{t-1}^i, X_t^{R_i}, C_t) \quad (10)$$

In other words, the weather state in region  $i$  today depends on the weather state in region  $i$  yesterday, the atmospheric data in region  $i$  today and the overall continental weather state today. The probability distribution of the continental weather state is completely defined by the continental atmospheric variables.

## SUMMARY

Simulations of precipitation conditional on a set of atmospheric data can be used to downscale the output of a GCM into local precipitation, and to assess the regional and local effects of climate change scenarios (by conditioning the simulations on the output of GCM runs assuming altered climate scenarios). This does, of course, require the untestable assumption that precipitation reacts to the altered climate in the same fashion as it does to the historical climate. In other words, we must assume that the climate change affects only the distribution of weather states, rather than which weather states are appropriate.

## REFERENCES

- Bardossy, A. and E. J. Plate (1992) Space-time models for daily rainfall using atmospheric circulation patterns. *Water Resour. Res.*, 28, 1247-1259.
- Hay, L., G. J. McCabe, D. M. Wolock, and M. A. Ayers (1991) Simulation of precipitation by weather type analysis. *Water Resour. Res.*, 27, 493-501.
- Hughes, J. P., D. P. Lettenmaier, P. Guttorp (1993) A stochastic approach for assessing the effects of changes in regional circulation patterns on local precipitation. *Water Resour. Res.*, 29, 3303-3315.
- Hughes, J. P. and P. Guttorp (1994a) A Class of Stochastic Models for Relating Synoptic Atmospheric Patterns to Regional Hydrologic Phenomena. *Water Resour. Res.*, 30, 1535-1546.
- Hughes, J. P. and P. Guttorp (1994b) Incorporating spatial dependence and atmospheric data in a model of precipitation. *J. Applied Meteor.*, 30, 1535-1546.

- McCabe, G. J., L. E. Hay, M. A. Ayers, and D. M. Wolock (1989) Assessment of climate change using weather type analysis. Presented at the National Conference on Hydraulic Engineering, Am. Soc. of Civ. Eng., New Orleans, La., Aug. 14-18.
- Wilson, L. L., D. P. Lettenmaier and E. Skyllingstad (1992) A hierarchical stochastic model of large-scale atmospheric circulation patterns and multiple station daily precipitation. *J. Geophys. Res.*, 97. 2791-2809.
- Zucchini, W. and P. Guttorp (1991) A hidden Markov model for space-time precipitation. *Water Resour. Res.*, 27, 1917-1923.

# Explicit Consideration of Multiple Landscape Scales While Selecting Spatial Resolutions

John B. Collins <sup>1</sup> and Curtis E. Woodcock <sup>2</sup>

---

**Abstract.** – The information content of a digital image – defined as the spatial variation of the image data – depends on data resolution. Spatial data may vary at a number of different scales simultaneously, so the choice of an appropriate resolution depends on the information content at the landscape scale relevant to a particular application. A useful framework for studying variation at different scales in digital images is provided by a nested hierarchical model of the landscape. The landscape is modeled as being composed of spatial units of a typical size, which are themselves composed of smaller units, and so on. The overall variation of the image data can be decomposed into variation due to effects of different levels of such a hierarchy. The hierarchical decomposition of the variation in a digital image is reflected by a similar hierarchical decomposition of the scene semivariance. Decomposed semivariograms are used to assess the relationships between data resolution and information content at particular landscape scales.

---

## INTRODUCTION

The issue of scale is of fundamental importance to disciplines concerned with phenomena which vary spatially. Yet the relationship between the scale at which a phenomenon is observed and that phenomenon's characteristic scale of variation is not completely understood. In this discussion, *scale of observation* is defined as the spatial interval over which observations are made. When data are collected by remote sensing devices, this term corresponds to sensor resolution, or size of the Instantaneous Field of View (IFOV). The *characteristic scale* of a phenomenon is the range of resolutions over which that phenomenon may be characterized. These definitions correspond to those proposed by Quattrochi (1993).

---

<sup>1</sup> Research Assistant, Boston University Department of Geography, Boston, MA.

<sup>2</sup> Associate Professor of Geography, Boston University, Boston, MA.

In remote sensing investigations, resolution is an important consideration. The information contained in an image, which can be identified as the overall variance of the image data, declines as resolution becomes more coarse. The simplest view of the issue is one which regards the selection of an appropriate resolution as a compromise between greater information content (fine resolutions) and more manageable data volume (coarse resolutions). But for some applications, greater spatial variation is not always desirable. For example, classification of remotely sensed imagery is most successful when pixel size is large enough that some amount of averaging of within-class variation occurs (Markham and Townshend, 1981).

The fact that variation of spatial data occurs at multiple scales has an explanation in the field of landscape ecology. It is becoming an accepted tenet of that field that landscapes are often arranged hierarchically (O'Neill et al., 1986). That is, landscapes are often composed of objects which are internally homogeneous and externally variable with respect to spatially varying processes. These objects may be composed of smaller objects, and so on. For example, a landscape may be composed of a number of watersheds, each containing a collection of vegetation stands, which each contain individual plants. Associated with each level of such a hierarchy is a set of processes acting at that level. The range of sizes of objects on a given level approximately defines the spatial frequencies at which those processes vary – i.e. their characteristic scales. So multiple scales of variation of spatial data arise from a scene's hierarchical structure.

The choice of an appropriate resolution for a particular investigation depends on prior knowledge of the characteristic scale of variation of the phenomenon of interest. This paper presents methods to help make such a choice under explicit consideration of the hierarchical nature of landscapes. A review of the use of hierarchical partitionings of spatial data is presented, followed by an explanation of geostatistical methods for investigating resolution-dependent characteristics of digital images.

## **SPATIAL HIERARCHIES**

A number of studies use a hierarchical model to identify multiple landscape scales. A common paradigm involves partitioning a scene into objects of regular size and shape – usually squares or rectangles – each of which are characterized by some aggregation of the underlying spatially distributed variable (Grieg-Smith, 1963; Townshend and Justice, 1995). Then there is a clear and explicit relationship between any level of the hierarchy and a scale of observation. By examining some attribute of each level of the hierarchy – especially the data variance – multiple scales of variation may be revealed.

An alternate approach is the use of hierarchies of irregular-shaped objects, such as political boundaries (Moellering and Tobler, 1972). The characteristic scales associated with any level are identified with the range of sizes of objects on that level. However, the hierarchies so constructed can be rather arbitrary.

Woodcock et al. (1994) describe a forest mapping project which makes use of a hierarchical model dividing the landscape into ecologically meaningful units.

(see also Woodcock and Harward, 1992; Collins et al., 1995). Under this model, the entire image is considered to be composed of a number of vegetation *classes* such as *conifer*, *hardwood*, and *brush*. Each of these can be broken down into smaller homogeneous patches, referred to as *regions*. Within the *conifer* class, for example, a region corresponds to a forest stand. Regions, in turn, are collections of individual plants. Such a hierarchy can be associated with ecosystem processes, which tend to be more homogeneous within units and heterogeneous between units at any level.

The partitioning of a scene into ecologically meaningful units prior to analysis constitutes explicit identification of the characteristic scales under study. Knowledge of these characteristic scales is important for assessing the relationship between resolution and information content for a given application.

### Statistical Analysis of Hierarchical Data

A general and mathematically precise definition of a hierarchical scene model is needed. Consider a regionalized variable  $Z$  which depends on vector location  $s$  within some domain set  $D_{(0)}$ . When discussing analysis of digital images,  $D_{(0)}$  is a subset of two-dimensional space  $R^2$ , and  $Z$  could be radiance in some wavelength band or any other spatially-varying quantity. The hierarchical model consists of partitioning the domain set  $D_{(0)}$  into a family of objects (sets)  $\{D_{(1)}\}$ , which form level 1 of a hierarchy. Each of the level-1 objects may be divided into a number of level-2 objects, which can collectively be referred to as  $\{D_{(2)}\}$ . Further partitions proceed in an obvious manner. If level  $n$  is the highest level of the hierarchy, then the family of domain sets  $\{D_{(n)}\}$  is the collection of infinitesimal points on which  $Z$  is defined. Any point  $s \in D_{(0)}$  will be contained in exactly one set associated with each level of the hierarchy. Such a set of level  $k$  is  $D_{(k)}(s)$ . This terminology is illustrated by Figure 1.

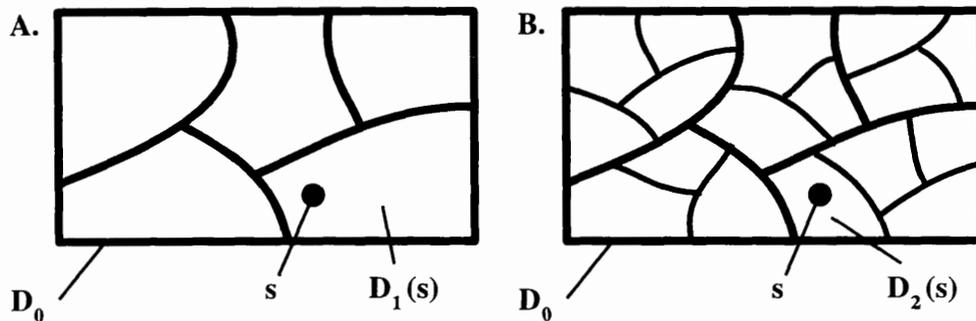


Figure 1. A hierarchical partitioning of a scene. A. The set of level-1 objects  $\{D_{(1)}\}$  forming a partition of  $D_{(0)}$ . B. The level-2 objects  $\{D_{(2)}\}$  further partitioning the level-1 objects.

Determination of the relative contributions of different levels of a hierarchy to overall variation in a data set is the goal of Hierarchical Analysis of Variance (Moellering and Tobler, 1972). Under the Hierarchical ANOVA model, the observed value at a point in the scene is considered equal to the overall mean for the area, plus a deviation, or *effect*, characteristic of each level of the hierarchy. The effect of level  $k$  of a hierarchy at point  $s$  is defined as  $\bar{D}_{(k)}(s) - \bar{D}_{(k-1)}(s)$ , where  $\bar{D}_{(k)}(s)$  is the mean value of  $Z$  over the level- $k$  object containing  $s$ . Note that each effect constitutes a regionalized variable in its own right. Further elaboration of the Hierarchical ANOVA model can be found in standard texts on the subject (e.g. Dunn and Clark, 1974). The model leads to the following important result.

$$\sigma_Z^2 = \sigma_{(1)}^2 + \sigma_{(2)}^2 + \cdots + \sigma_{(n)}^2 \quad , \quad (1)$$

where  $\sigma_Z^2$  represents the overall variance of  $Z$ , and  $\sigma_{(i)}^2$  represents the variance of the effect of level  $i$ . The overall data variance decomposes into the sum of variances of each effect. Unfortunately, the ability to make statistical inferences from a hierarchical partitioning of variance relies on the assumption of independence of observations. Since this assumption is badly violated for most spatial data sets, the technique is suited only for descriptive purposes.

When the regionalized variable  $Z$  exhibits appropriate stationarity characteristics, its variation is described by the scene semivariogram (Isaaks and Srivastava, 1989; Cressie, 1993).

$$\gamma_Z(h) = \frac{1}{2} E [Z(s) - Z(s+h)] \quad , \quad (2)$$

where  $h$  is a vector lag. Similarly, relationships between two regionalized variables  $Y$  and  $Z$  are described by the cross-semivariogram.

$$\gamma_{YZ}(h) = \frac{1}{2} E \{ [Z(s) - Z(s+h)] [Y(s) - Y(s+h)] \} \quad . \quad (3)$$

The parameters describing the scene semivariogram – its range, sill, and derivative at the origin – can be shown to be related to quantifiable attributes of a scene (Woodcock et al. 1988a+b).

Since the semivariogram provides a better characterization of the variation of a regionalized variable than does the variance, it is useful to ask whether there is a geostatistical analogy to Eq. 1. Collins et al. (1995) present an example to show that under the Hierarchical ANOVA model, the scene semivariogram decomposes as

$$\gamma_Z(h) = \gamma_{(1)}(h) + \cdots + \gamma_{(n)}(h) + 2 \sum_{j=1}^n \sum_{k=1}^{j-1} \gamma_{(j)(k)}(h) \quad , \quad (4)$$

where  $\gamma_Z$  represents the overall scene semivariogram,  $\gamma_{(i)}$  represents the semivariogram associated with the effect of level  $i$  of the hierarchy, and  $\gamma_{(j)(k)}$  is the cross-semivariogram between the effects of levels  $j$  and  $k$ . It is interesting to note that when overall scene variance is considered (Eq. 1) all cross-product terms vanish, but that this is not the case for spatial variance. The cross-semivariograms in (4) arise as the result of fine-scale continuity across

boundaries of larger-scale objects (Collins and Woodcock, manuscript in preparation). They generally have small magnitudes and relatively short ranges, and approach zero at large lags. The relevance of Eq. 4 to the present study is its indication that semivariograms associated with particular levels of a hierarchy – i.e. with different characteristic scales in the landscape – may be discussed individually.

## EFFECT OF RESOLUTION ON SPATIAL STRUCTURE

The scale-specific semivariograms included in Eq. 4 may be used to retrieve quantitative information about scene variation at arbitrary resolutions. In geostatistical parlance, sensor resolution – the size and shape of the IFOV – is called the *support* of the measurements. There exists a well-developed theory of the effect of changes of support on regionalized variables (Rendu, 1978; Zhang et al., 1990), and its applications in the context of remote sensing (Jupp et al., 1988+1989; Atkinson and Curran, 1995). This section discusses some of the ways in which information relating to scale of observation can be derived from semivariograms. Notationally, subscripts indicative of levels in a hierarchy are dropped, as the ideas apply to any regionalized variable with finite variance. The relevance to multiple scales of variation comes from the indication of Eq. 4 that these methods can be applied to scale-specific semivariograms.

### Regularization and Deregularization

As was noted above, the effect of a level in a hierarchy is a regionalized variable, and is defined at all points in a scene. However, point-scale observations are never available from a remote sensing device, since there is always some averaging of scene radiance over the sensor IFOV. That is, the data and the semivariograms derived from it are *regularized*. It is desirable to derive the point-scale, or *punctual* semivariograms from the regularized ones.

In the theoretical development which follows, the support of observations is modeled as a set in the domain space  $D_{(0)}$  over which the continuously-varying values of some effect are averaged. A useful quantity is the average punctual semivariance between two sets in the scene, denoted  $A$  and  $B$ .

$$\bar{\gamma}(A, B) = \frac{1}{\text{Mes}(A)\text{Mes}(B)} \int_A \int_B \gamma(z-z') dz' dz \quad (5)$$

$\text{Mes}(\ast)$  denotes the Lebesgue measure of a set (i.e. its area in this context). This definition is particularly useful when the sets  $A$  and  $B$  are shifts of a sensor IFOV. Specifically, let  $w$  denote some support, and let  $w_h$  denote its translation by distance  $h$ . The regularized semivariogram at a lag of  $h$  can be found from the punctual semivariogram by (Rendu, 1978)

$$\gamma_w(h) = \bar{\gamma}(w; w_h) - \bar{\gamma}(w, w) \quad , \quad (6)$$

This relationship cannot be inverted to find the punctual semivariogram from

its regularized counterpart. But Atkinson and Curran (1995) present a numerical technique based on this relationship which may be used to estimate it.

The effects of regularization are most pronounced on higher levels of a scene hierarchy, where the characteristic scales of variation are the smallest. When scales of variability are much smaller than the support, 'microstructures' in the data are undetected and spatial variation is lost. This phenomenon plus the effect of measurement error can possibly lead to an observed 'nugget effect' or a non-zero intercept of a regularized semivariogram. At scales larger than that at which data are collected, the nugget effect shows no spatial structure. This leads to a simple relationship between the nugget effect observed via support  $w$  ( $N_w$ ) and that observed at a larger support  $W$  ( $N_W$ ):

$$N_W = \frac{\text{Mes}(w)}{\text{Mes}(W)} N_w \quad . \quad (7)$$

The nugget effect behaves differently under changes of support than does the spatially correlated component of variation. Thus it is useful to express the regularized semivariogram as (Zhang et al., 1990):

$$\gamma_w(h) = N_w + \gamma_w^*(h) \quad , \quad (8)$$

where  $\gamma_w^*$  is a regularized zero-nugget semivariogram, from which a punctual zero-nugget semivariogram  $\gamma^*$  may be estimated. Then using 6, 7, and 8 the semivariogram for any support  $W$  can be expressed as

$$\gamma_W(h) = \frac{\text{Mes}(w)}{\text{Mes}(W)} N_w + \bar{\gamma}^*(W, W_h) - \bar{\gamma}^*(W, W) \quad . \quad (9)$$

Since parameters describing semivariograms relate to properties of the scene under study, the ability to determine semivariograms for arbitrary supports is useful. When used with a hierarchical decomposition of a scene semivariogram, it can yield useful information concerning the relationship between image resolution and spatial variation at different characteristic scales.

### Determination of Variance at Arbitrary Resolutions

As was mentioned above, the overall scene variance is perhaps the best measure of its information content. As a rule, the variance decreases as the size of support increases. But the precise rate of decrease depends on the autocorrelation structure of the image, as revealed by the semivariogram. If autocorrelation between points is strong, then regularization will have little effect since averaging over the IFOV will tend to occur between points with similar values.

This assertion is justified mathematically. Let  $D_{(0)}$  represent an area being studied. If this area has been sampled on supports  $w$ , then the variance of supports  $W$  within  $D_{(0)}$  is (Zhang et al., 1990)

$$V(W, D_{(0)}) = \left[ \frac{\text{Mes}(w)}{\text{Mes}(W)} - \frac{\text{Mes}(w)}{\text{Mes}(D_{(0)})} \right] N_w + \bar{\gamma}^*(D_{(0)}, D_{(0)}) - \bar{\gamma}^*(W, W) \quad (10)$$

The first term accounts for the nugget effect among the supports  $W$ , and the last two terms deal with the spatially-correlated component of variation. Eq. 10 gives the variance for arbitrary resolutions, and so is useful for a number of purposes.

## CONCLUSION

Methods to determine the effect of sensor resolution on different landscape scales have been presented. Characteristic scales of variation can be identified by partitioning a scene hierarchically into ecologically meaningful units. In order not to restrict the discussion to a particular classification of scene objects, a general hierarchical model has been used. Use of different landscape hierarchies may be called for in different situations, but the ideas discussed here continue to apply.

Semivariograms may be calculated characterizing the spatial variation at each landscape scale, and may be used to derive information about scene structure at arbitrary resolutions. Throughout the discussion above, the assumption has been made that the data are second-order stationary at all scales, a requirement for the calculation of semivariograms. This assumption needs to be justified in order for the analysis of semivariance to be appropriate. If this condition is met, the relationships between resolution and semivariance presented here are general with respect to directional qualities of variation. All results apply equally for isotropic and non-isotropic semivariograms.

## REFERENCES

- Atkinson, P. M. and Curran, P. J. (1995). Defining an optimal size of support for remote sensing investigations, *IEEE Transactions on Geoscience and Remote Sensing*, 33(3):768-776.
- Collins, J. B., Woodcock, C. E. and Jupp, D. L. B. (1995). Spatial dependence and nested hierarchical scene models, *1995 ACSM/ASPRS Annual Convention and Exposition Technical Papers*: 535-544.
- Cressie, N. A. C. (1993). *Statistics for Spatial Data*, John Wiley and Sons, Inc., New York.
- Dunn, O. J. and Clark, V. A. (1974). *Applied Statistics: Analysis of Variance and Regression*, John Wiley and Sons, Inc., New York.
- Hobbs, R. J. and Mooney, H. A., Eds. (1990). *Remote Sensing of Biosphere Functioning*, Springer-Verlag, New York.
- Isaaks, E. H. and Srivastava, R. M. (1989). *An Introduction to Applied Geostatistics*, Oxford University Press, Oxford.
- Jupp, D. L. B., Strahler, A. H., and Woodcock, C. E. (1988). Autocorrelation and regularization in digital images I: basic theory, *IEEE Transactions on Geoscience and Remote Sensing*, 26(4):463-473.

- Jupp, D. L. B., Strahler, A. H., and Woodcock, C. E. (1989). Autocorrelation and regularization in digital images II: simple image models, *IEEE Transactions on Geoscience and Remote Sensing*, 27(3):247-258.
- Markham, B. L., and Townshend, J. R. G. (1981). Land cover classification accuracy as a function of sensor spatial resolution, *Proceedings of the 15th Int. Symp. on Remote Sensing of Environment*, Ann Arbor, Mi.: 1075-1090.
- Matheron, G. (1963). Principles of geostatistics, *Economic Geology*, 58:1246-1266.
- Mollering, H. and W. Tobler (1972). Geographical variances, *Geographical Analysis*, 4 (1):35-50.
- O'Niell, R. V., DeAngelis, D. L., Waide, J. B., and Allen, T. F. H. (1986). *A Hierarchical Concept of Ecosystems*, Princeton University Press, Princeton, New Jersey.
- Quattrochi, D. A. (1993). The need for a lexicon of scale terms in integrating remote sensing data with geographic information systems, *Journal of Geography*, 92(3):206-212.
- Rendu, J. M. (1978). *An Introduction to Geostatistical Methods of Mineral Evaluation*, South African Institute of Mining and Metallurgy, Johannesburg.
- Townshend, J. R. G. and Justice, C. O. (1995). Spatial variability of images and the monitoring of changes in the normalized difference vegetation index, *International Journal of Remote Sensing*, 16(12):2187-2195.
- Woodcock, C.E., Collins, J., Gopal, S., Jakabhazy, V., Li, X., Macomber, S., Ryherd, S., Wu, Y., Harward, V. J., Levitan, J., and Warbington, R. (1994). Mapping Forest Vegetation Using Landsat TM Imagery and a Canopy Reflectance Model, *Remote Sensing of Environment*, 50:240-254.
- Woodcock, J. E. and J. Harward (1992). Nested-Hierarchical Scene Models and Image Segmentation, *International Journal of Remote Sensing*, 13(16):3167-3187.
- Woodcock, C. E., Strahler, A. H., and Jupp, D. L. B. (1988). The use of variograms in remote sensing: I. scene models and simulated images, *Remote Sensing of Environment*, 25:323-348.
- Woodcock, C. E., Strahler, A. H., and Jupp, D. L. B. (1988). The use of variograms in remote sensing: II. real digital images, *Remote Sensing of Environment*, 25:349-379.
- Zhang, R., Warrick, A. W., and Meyers, D. E. (1990). Variance as a function of sample support size, *Mathematical Geology* 22(1):107-121.

# Using Trend Surface Methodology to Compare Spatial Surfaces

Dale Zimmerman<sup>1</sup>, Zhi-Jun Liu<sup>2</sup>, and George Hallberg<sup>3</sup>

---

Trend surface analysis is a very popular method for describing spatial data. When data on more than one response variable are available, it is generally desirable to compare the corresponding fitted surfaces obtained by trend surface analysis. Existing methods for such comparisons have serious deficiencies. The objective of this article is to develop statistically valid techniques for such comparisons which retain the familiarity and relative simplicity of trend surface analysis. In particular, we develop tests for surface parallelism and surface coincidence. The methods are illustrated by applying them to groundwater quality data from an aquifer in northeast Iowa.

---

## INTRODUCTION

The multiple linear regression method known as polynomial trend surface analysis (TSA) is used widely in the physical sciences to describe geostatistical data. TSA is based on an assumption that the available data can be usefully described by a model in which each observation is equal to the sum of a trend component, generally taken to be a polynomial function of the spatial coordinates, and an error component. For example, for data taken from a two-dimensional region, a  $p$ th-order TSA model is given by

$$Z(x, y) = \beta_{00} + \beta_{10}x + \beta_{01}y + \beta_{20}x^2 + \beta_{11}xy + \beta_{02}y^2 + \cdots + \beta_{rs}x^r y^s + \epsilon(x, y).$$

Here,  $Z(x, y)$  is the variate of interest's value at the site with Cartesian coordinates  $(x, y)$ ;  $\beta_{00}, \beta_{10}, \dots, \beta_{rs}$  are unknown parameters;  $r + s = p$ , the order of the polynomial; and  $\epsilon(x, y)$  is the error component. In the earliest applications of TSA, inference on model parameters was based on an ordinary least squares fit and thus the errors were taken, either explicitly or implicitly, to be independent normally distributed random variables with zero mean and constant variance (Grant 1957; Krumbein 1959). Standard  $F$ -tests from multiple regression analysis were used to determine the order of the trend surface. Later, however, it was argued that in many situations small-scale positive spatial dependence is likely to exist among the errors, rendering the ordinary least

---

<sup>1</sup>Associate Professor of Statistics, University of Iowa, Iowa City, IA.

<sup>2</sup>Research Associate, University Hygienic Laboratory, Iowa City, IA.

<sup>3</sup>Geologist and Associate Director, University Hygienic Laboratory, Iowa City, IA.

squares estimates of model coefficients inefficient, biasing tests on the parameters, and tending to result in a fitted surface of too high an order (Cliff and Ord 1973; Ripley 1981). More recently, Agterberg (1984) and Haining (1987) have used generalized least squares to fit TSA models with spatially correlated errors.

Although some geostatistical problems in which TSA can be used involve fitting a surface to data on a single response variable, others involve fitting surfaces to data on several response variables. If these variables are commensurate or correspond to a single variable measured on several occasions, it may be of interest to compare the surfaces to each other. For example, one may want to compare the surface that describes current levels of a groundwater contaminant in an aquifer to the surface that corresponds to levels determined 10 years previously. The objective of this paper is to extend the statistical methodology of TSA to handle such comparisons.

The comparison of two or more fitted polynomial trend surfaces is not a new problem; indeed, it was an active area of research in the 1950s and 1960s as TSA was becoming popular among mathematical geologists and geographers. Three main approaches to the problem were proposed: a point correlation approach, a regression coefficient comparison, and a gradient-based approach. For a review of these, see Goodman (1983). Although these approaches are popular, they have several shortcomings, not the least of which is that none of them yield a statistically valid test for surface similarity (Zimmerman et al. 1996).

The methodology for comparing spatial surfaces that we propose here does not suffer from these shortcomings. In particular, it permits one to formally and validly test for coincidence and parallelism of the surfaces in a manner which accounts for spatial correlation.

## METHODOLOGY

### Preliminaries

To simplify the exposition, we consider an important particular case of the general problem. We assume that only two surfaces, with common spatial domain  $D$ , are to be compared. The surfaces may correspond either to two commensurate variables or to the same variable measured on two occasions; the terminology we use, however, presumes the latter. We also assume that  $D \subset R^2$  and denote an arbitrary location in  $D$  as  $\mathbf{s} = (x, y)'$ . The cases of more than two surfaces or more than two dimensions can be handled by the same general methods to be presented here, but the notation becomes overly cumbersome.

The basis of our approach for comparing surfaces is to regard the data from which each fitted surface is constructed as a partial realization of a random process. Let  $\{Z_1(\mathbf{s}): \mathbf{s} \in D\}$  and  $\{Z_2(\mathbf{s}): \mathbf{s} \in D\}$  denote the random processes

corresponding to the variable of interest (or to perhaps a transformation, such as logarithms, of the variable of interest) on the two occasions. It is assumed that these processes are governed by the model

$$Z_k(\mathbf{s}) = \sum_{1 \leq i+j \leq p} \beta_{ijk} x^i y^j + \epsilon_k(\mathbf{s}) \quad (1)$$

where  $\{\beta_{ijk}: 0 \leq i+j \leq p; k=1,2\}$  are unknown parameters and  $\{\epsilon_1(\mathbf{s})\}$  and  $\{\epsilon_2(\mathbf{s})\}$  are independent, stationary, zero-mean Gaussian processes with covariance functions  $C_1(\mathbf{h}) \equiv \text{cov}\{\epsilon_1(\mathbf{s}), \epsilon_1(\mathbf{s}+\mathbf{h})\}$  and  $C_2(\mathbf{h}) \equiv \text{cov}\{\epsilon_2(\mathbf{s}), \epsilon_2(\mathbf{s}+\mathbf{h})\}$ , respectively. We assume initially that  $p$  and the covariance functions are known; subsequently we will discuss their estimation.

Let  $\mathbf{s}_1 = (x_1, y_1)'$ ,  $\dots$ ,  $\mathbf{s}_{n_1} = (x_{n_1}, y_{n_1})'$  denote the  $n_1$  locations where  $\{Z_1(\mathbf{s})\}$  is observed, and let  $\mathbf{s}_{n_1+1} = (x_{n_1+1}, y_{n_1+1})'$ ,  $\dots$ ,  $\mathbf{s}_{n_1+n_2} = (x_{n_1+n_2}, y_{n_1+n_2})'$  denote the  $n_2$  locations where  $\{Z_2(\mathbf{s})\}$  is observed. Regarding the relationship of the first set of locations to the second set, two cases will be considered: (1) arbitrary sampling networks, in which no relationship between the first set and second set is specified, and (2) identical sampling networks, in which  $n_1 = n_2 \equiv n$  and  $\mathbf{s}_l = \mathbf{s}_{n+l}$  for  $l = 1, \dots, n$  (perhaps after a suitable permutation of the locations' labels). The second case permits an analysis based on paired differences among observations at the same locations and thus bears a resemblance to the case of paired-difference data in a classical two-sample problem. If, for fixed  $\mathbf{s}$ , the errors  $\epsilon_1(\mathbf{s})$  and  $\epsilon_2(\mathbf{s})$  actually exhibit positive temporal dependence rather than being independent as assumed in model (1), then the variance of the difference in these two errors will be smaller than the sum of their two variances. This reduction in variation, if large enough and consistent across locations, can result in substantially more powerful tests for surface coincidence and parallelism.

Finally, let  $\mathbf{Z}_1 = [Z_1(\mathbf{s}_{11}), \dots, Z_1(\mathbf{s}_{1n_1})]'$  and  $\mathbf{Z}_2 = [Z_2(\mathbf{s}_{21}), \dots, Z_2(\mathbf{s}_{2n_2})]'$  denote the vectors of responses observed at the two times, and let  $\boldsymbol{\epsilon}_1 = [\epsilon_1(\mathbf{s}_{11}), \dots, \epsilon_1(\mathbf{s}_{1n_1})]'$  and  $\boldsymbol{\epsilon}_2 = [\epsilon_2(\mathbf{s}_{21}), \dots, \epsilon_2(\mathbf{s}_{2n_2})]'$  denote the corresponding vectors of errors. Put  $\mathbf{Z} = [\mathbf{Z}'_1, \mathbf{Z}'_2]'$  and  $\boldsymbol{\epsilon} = [\boldsymbol{\epsilon}'_1, \boldsymbol{\epsilon}'_2]'$ .

### Arbitrary Sampling Networks

Under the stated assumptions and using the notation defined previously, the model for the observed data can be written as

$$Z_l = \begin{cases} \beta_{001} + \sum_{1 \leq i+j \leq p} \beta_{ij1} x_i^i y_l^j + \epsilon_l & \text{for } l = 1, \dots, n_1 \\ \beta_{002} + \sum_{1 \leq i+j \leq p} \beta_{ij2} x_i^i y_l^j + \epsilon_l & \text{for } l = n_1 + 1, \dots, n_1 + n_2 \end{cases} \quad (2)$$

where  $E(\boldsymbol{\epsilon}) = \mathbf{0}$  and  $\text{var}(\boldsymbol{\epsilon}) \equiv \text{diag}(\boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{22}) \equiv \boldsymbol{\Sigma}$ , which we take to be positive definite. Here,  $\boldsymbol{\Sigma}_{kk}$  is the  $n_k \times n_k$  matrix with  $(r, s)$ th element given by  $C_k(\mathbf{s}_{kr} - \mathbf{s}_{ks})$ . Let  $q = (p+1)(p+2)/2$  and let  $\mathbf{X}_1$  denote the  $(n_1 + n_2) \times 2q$  model matrix corresponding to model (2), which is assumed to have full column

rank, and let  $\boldsymbol{\beta}$  denote the associated  $2q \times 1$  vector of coefficients. The best linear unbiased estimator of  $\boldsymbol{\beta}$  is given by  $\hat{\boldsymbol{\beta}} = (\mathbf{X}'_1 \boldsymbol{\Sigma}^{-1} \mathbf{X}_1)^{-1} \mathbf{X}'_1 \boldsymbol{\Sigma}^{-1} \mathbf{Z}$ .

For purposes of testing for parallelism or coincidence of the two surfaces in model (2), two submodels are defined as follows.

*Parallelism Model*

$$Z_l = \begin{cases} \beta_{001} + \sum_{1 \leq i+j \leq p} \beta_{ij} x_i^i y_l^j + \epsilon_l & \text{for } l = 1, \dots, n_1 \\ \beta_{002} + \sum_{1 \leq i+j \leq p} \beta_{ij} x_i^i y_l^j + \epsilon_l & \text{for } l = n_1 + 1, \dots, n_1 + n_2 \end{cases}$$

*Coincidence Model*

$$Z_l = \beta_{00} + \sum_{1 \leq i+j \leq p} \beta_{ij} x_i^i y_l^j + \epsilon_l \text{ for } l = 1, \dots, n_1 + n_2.$$

Now let  $RSS_m = \mathbf{Z}' \boldsymbol{\Sigma}^{-1} \mathbf{Z} - \mathbf{Z}' \boldsymbol{\Sigma}^{-1} \mathbf{X}_m (\mathbf{X}'_m \boldsymbol{\Sigma}^{-1} \mathbf{X}_m)^{-1} \mathbf{X}'_m \boldsymbol{\Sigma}^{-1} \mathbf{Z}$  for  $m = 1, 2, 3$ , where  $\mathbf{X}_2$  and  $\mathbf{X}_3$  are the model matrices corresponding to the parallelism and coincidence models, respectively. Then, the parallelism hypothesis can be tested by comparing the value of

$$F_P = \frac{(RSS_2 - RSS_1)/(q-1)}{RSS_1/(n_1 + n_2 - 2q)}$$

to a suitable percentage point of a (central)  $F$  distribution with  $q-1$  and  $n_1 + n_2 - 2q$  degrees of freedom. Similarly, we can test for coincidence by comparing the value of

$$F_C = \frac{(RSS_3 - RSS_1)/q}{RSS_1/(n_1 + n_2 - 2q)}$$

to a suitable percentage point of an  $F$  distribution with  $q$  and  $n_1 + n_2 - 2q$  degrees of freedom.

The procedures just described are identical to the classical tests for parallelism or coincidence of regression surfaces described, for example, by Weisberg (1985) except that, to account for spatial correlation, the residual sums of squares used here are based on generalized least squares (GLS), rather than ordinary least squares (OLS), fits of the models. It should be noted that a "spatial analysis of variance," of which the analyses of variance that correspond to these procedures are examples, has been used before for problems other than testing for parallelism or coincidence of surfaces; see Griffith (1978) and Gotway and Cressie (1990).

### Identical Sampling Networks

Now suppose that  $n_1 = n_2 \equiv n$  and  $\mathbf{s}_l = \mathbf{s}_{n+l}$  for  $l = 1, \dots, n$ . Define the paired differences

$$\begin{aligned} D_l &\equiv Z_2(\mathbf{s}_l) - Z_1(\mathbf{s}_l) = Z_{n+l} - Z_l, \\ \delta_l &\equiv \epsilon_2(\mathbf{s}_l) - \epsilon_1(\mathbf{s}_l) = \epsilon_{n+l} - \epsilon_l \text{ for } l = 1, \dots, n \end{aligned}$$

and let  $\mathbf{D} = (D_l)$  and  $\boldsymbol{\delta} = (\delta_l)$ . It follows from model (2) that the model for  $D_l$  is

$$D_l = \gamma_{00} + \sum_{1 \leq i+j \leq p} \gamma_{ij} x_i^i y_l^j + \delta_l \text{ for } l = 1, \dots, n \quad (3)$$

where  $\gamma_{ij} = \beta_{ij2} - \beta_{ij1}$ ,  $E(\boldsymbol{\delta}) = \mathbf{0}$ , and  $\text{var}(\boldsymbol{\delta}) = \boldsymbol{\Sigma}_{11} + \boldsymbol{\Sigma}_{22} \equiv \boldsymbol{\Sigma}_{\delta}$ . Let  $\mathbf{W}$  denote the  $n \times q$  model matrix corresponding to model (3) and let  $\boldsymbol{\gamma}$  denote the associated  $q \times 1$  vector of coefficients. Based on this model, the best linear unbiased estimator of  $\boldsymbol{\gamma}$  is given by  $\hat{\boldsymbol{\gamma}} = (\mathbf{W}'\boldsymbol{\Sigma}_{\delta}^{-1}\mathbf{W})^{-1}\mathbf{W}'\boldsymbol{\Sigma}_{\delta}^{-1}\mathbf{D}$ .

Testing for parallelism corresponds, in this situation, to testing that all the coefficients in model (3) except  $\gamma_{00}$  equal 0 and hence can be tested by the mean-corrected overall  $F$ -test from a GLS analysis. That is, we compare the value of

$$F_{P'} = \frac{\text{RegSS}_4/(q-1)}{\text{RSS}_4/(n-q)}$$

to a suitable percentage point of an  $F$  distribution with  $q-1$  and  $n-q$  degrees of freedom, where  $\text{RegSS}_4 = \mathbf{D}'\boldsymbol{\Sigma}_{\delta}^{-1}\mathbf{W}(\mathbf{W}'\boldsymbol{\Sigma}_{\delta}^{-1}\mathbf{W})^{-1}\mathbf{W}'\boldsymbol{\Sigma}_{\delta}^{-1}\mathbf{D} - \mathbf{D}'\boldsymbol{\Sigma}_{\delta}^{-1}\mathbf{1}_n(\mathbf{1}_n'\boldsymbol{\Sigma}_{\delta}^{-1}\mathbf{1}_n)^{-1}\mathbf{1}_n'\boldsymbol{\Sigma}_{\delta}^{-1}\mathbf{D}$  and  $\text{RSS}_4 = \mathbf{D}'\boldsymbol{\Sigma}_{\delta}^{-1}\mathbf{D} - \mathbf{D}'\boldsymbol{\Sigma}_{\delta}^{-1}\mathbf{W}(\mathbf{W}'\boldsymbol{\Sigma}_{\delta}^{-1}\mathbf{W})^{-1}\mathbf{W}'\boldsymbol{\Sigma}_{\delta}^{-1}\mathbf{D}$ . Testing for coincidence corresponds in this situation to testing that *all* coefficients in model (3) equal 0 and hence can be tested by comparing

$$F_{C'} = \frac{\text{RegSS}_5/q}{\text{RSS}_4/(n-q)}$$

to a suitable percentage point of an  $F$  distribution with  $q$  and  $n-q$  degrees of freedom, where  $\text{RegSS}_5 = \mathbf{D}'\boldsymbol{\Sigma}_{\delta}^{-1}\mathbf{W}(\mathbf{W}'\boldsymbol{\Sigma}_{\delta}^{-1}\mathbf{W})^{-1}\mathbf{W}'\boldsymbol{\Sigma}_{\delta}^{-1}\mathbf{D}$ .

### Covariance Estimation and Order Determination

In practice, of course, the covariance matrices  $\boldsymbol{\Sigma}_{11}$  and  $\boldsymbol{\Sigma}_{22}$  (or  $\boldsymbol{\Sigma}_{\delta}$ , if a difference-based approach is possible) and the orders  $p_1$  and  $p_2$  of the two surfaces are unknown. The covariance matrices can be estimated using a classical geostatistical approach based on the sample semivariogram; see Cressie (1991, ch. 2). Once  $\boldsymbol{\Sigma}$  (or  $\boldsymbol{\Sigma}_{\delta}$ ) has been estimated, the tests for parallelism and coincidence that we have described can be carried out by substituting  $\hat{\boldsymbol{\Sigma}}$  for  $\boldsymbol{\Sigma}$  in the previously given expressions for the test statistics. Of course, the resulting test statistics no longer have  $F$  distributions, but provided that  $\hat{\boldsymbol{\Sigma}}$  is a consistent estimator of  $\boldsymbol{\Sigma}$ , the asymptotic distributions of the test statistics are  $F$  distributions and hence the tests are asymptotically valid; see Cressie (1991, pp. 242-243).

As for the determination of the order(s) of the surfaces, a stepwise procedure is described in Zimmerman et al. (1996) in which, starting with  $p_0 = 0$ , a  $(p_0 + 1)$ -order model is fitted by OLS; its residuals are tested for spatial correlation; and then an  $F$ -test (based on either an OLS or GLS fit) comparing the  $p_0$ -order and  $(p_0 + 1)$ -order models is carried out.

## EXAMPLE: NITRATE CONCENTRATION IN BIG SPRING BASIN, IOWA

Big Spring Basin is a groundwater basin located in Clayton County in northeast Iowa. It is named after Big Spring, the largest spring in the state, which discharges from the underlying Galena aquifer. The area of the basin is 267 square km (103 square miles). Because of the proximity of the aquifer to the land surface and the existence of karst features (such as sinkholes) at the surface, the Galena aquifer in this basin is highly responsive to recharge events and susceptible to a relatively rapid influx of contaminants. The land in the basin is essentially all used for agriculture. There are no significant urban or industrial areas, landfills, or other major point sources that may affect groundwater quality. The combination of pure agricultural use and a responsive hydrogeological system in the basin offers a good opportunity for studying the impact of agricultural non-point source pollution on groundwater quality.

The Geological Survey Bureau of the Iowa Department of Natural Resources conducted a basin-wide sampling of private wells in the Galena aquifer in 1989, which was a drought year, and 1992, which was a relatively wet year. Wells were selected somewhat haphazardly, and inclusion in the sample depended on the presence and cooperation of the property owner, among other factors. Complete data were available from both years for 125 wells. All wells were sampled during a period of 7-10 days in October of the respective year so that the effects of short-term climatic and hydrologic events, as well as seasonal variability, would be minimized. Concentrations of several contaminants in the samples were analyzed by the University Hygienic Laboratory in Iowa City, a U.S. Environmental Protection Agency certified contract laboratory. Here we consider only the nitrate ( $\text{NO}_3^-$ ) concentrations, which were measured in mg/L and transformed to the log scale. In addition to the nitrate concentrations, the depth and  $(x, y)$  coordinates of each well was recorded. All three coordinates were scaled to lie in the range  $[-1, 1]$ .

One of the objectives of the investigation was to determine whether the surface of nitrate concentrations over the basin was consistent with what would be expected from scientific considerations. For example, it is well known that nitrate concentrations decrease with increasing depth. Moreover, in the eastern half of the basin there were a large number of sinkholes, so higher concentrations were expected there. Another objective of the investigation was to determine how the nitrate concentration surface had changed from 1989 to 1992. In particular, since precipitation levels differed substantially in the two years, it was hoped that changes in the surface could lead to the formation of tentative hypotheses on how precipitation amounts and responsiveness of the basin may interact to affect groundwater nitrate concentrations.

To address these objectives, first we performed a trend surface analysis

separately for each year. Application of our order determination procedure resulted in a first-order model for both 1989 and 1992. Examination of the sample semivariograms of residuals, as well as Schmoyer's (1994) permutation test for spatial correlation, indicated that the spatial correlation was quite weak at the scale represented by most lags in the data set. Consequently, we fit first-order models by OLS, obtaining the fitted models

$$\hat{Z}_1(d, x, y) = \frac{1.66}{(0.27)} + \frac{2.22x}{(0.55)} - \frac{0.79y}{(0.42)} - \frac{1.45d}{(0.54)}$$

for 1989 and

$$\hat{Z}_2(d, x, y) = \frac{2.52}{(0.31)} + \frac{1.90x}{(0.63)} - \frac{0.45y}{(0.48)} - \frac{1.72d}{(0.61)}$$

for 1992. (Here,  $d$  represents depth and quantities in parentheses under the coefficients indicate their standard errors.) The signs and magnitudes of the estimated coefficients were consistent with what was expected from scientific considerations: (1) nitrate concentration decreased with increasing depth; (2) nitrate concentration increased as one moved from west to east; and (3) the relationship between nitrate concentration and the north-south coordinate was not as strong as the others.

Next, we compared the surfaces by testing for parallelism and coincidence using the  $F$ -tests we have previously described. The parallelism hypothesis was accepted ( $P = .95$ ) but the coincidence hypothesis, assuming parallelism, was strongly rejected ( $P < .0001$ ). This indicates that differences in precipitation levels (and possibly differences in other factors confounded with year) significantly affected the height, but not the shape, of the nitrate concentration surface.

The entire testing procedure was repeated using GLS fits of the two surfaces. The classical geostatistical procedure was used to estimate the covariance matrices for these fits. Results of the tests were qualitatively very similar to the results obtained under the assumption of no spatial correlation.

One interesting feature of the OLS analysis, however, was that the residuals from 1989 were significantly, positively correlated with the residuals (at the same sites) from 1992 ( $r = 0.70, P < .0001$ ). This suggests that although it may be reasonable to assume that  $\Sigma_{11}$  and  $\Sigma_{22}$  are scalar multiples of an identity matrix for these data, it may not be reasonable to assume that  $\Sigma$  is block diagonal, as in model (2). Furthermore, there was evidence of somewhat greater variation among the measurements in 1992 than in 1989. For these reasons, and since the data lie on identical sampling networks, we carried out the difference-based approach as well. The conclusions drawn from applying this approach were not fundamentally different than those obtained using the original (i.e. undifferenced) observations.

## REFERENCES

- Agterberg, F.P. 1984. Trend surface analysis. Pages 147-171 in *Spatial Statistics and Models*, eds. G.L. Gaile and C.J. Willmott, Dordrecht: D. Reidel.
- Cliff, A. and Ord, J. 1973. *Spatial Autocorrelation*. London: Pion.
- Cressie, N. 1991. *Statistics for Spatial Data*. New York: Wiley.
- Gotway, C.A. and Cressie, N.A.C. 1990. A spatial analysis of variance applied to soil-water infiltration. *Water Resources Research* 26, 2695-2703.
- Goodman, A. 1983. COMPARE: A Fortran IV program for the quantitative comparison of polynomial trend surfaces. *Computers & Geosciences* 9, 417-423.
- Grant, F.A. 1957. A problem in the analysis of geophysical data. *Geophysics* 22, 309-344.
- Griffith, D.A. 1978. A spatially adjusted ANOVA model. *Geographical Analysis* 10, 296-301.
- Haining, R. 1987. Trend-surface models with regional and local scales of variation with an application to aerial survey data. *Technometrics* 29, 461-469.
- Krumbein, W.C. 1959. Trend surface analysis of contour-type maps with irregular control-point spacing. *Journal of Geophysical Research* 68, 5869-5878.
- Ripley, B.D. 1981. *Spatial Statistics*. New York: Wiley.
- Schmoyer, R.L. 1994. Permutation tests for correlation in regression errors. *Journal of the American Statistical Association* 89, 1507-1516.
- Weisberg, S. 1985. *Applied Linear Regression*. New York: Wiley.
- Zimmerman, D.L., Liu, Z. and Hallberg, G. 1996. Using trend surface methodology and locally weighted regression to compare spatial surfaces. Unpublished Technical Report, Department of Statistics & Actuarial Science, University of Iowa.

## BIOGRAPHICAL SKETCH

Dale Zimmerman is an Associate Professor in the Department of Statistics and Actuarial Science at the University of Iowa. He graduated from Iowa State University in 1986 with a Ph.D. in Statistics. Dale teaches and conducts research in spatial statistics, environmental statistics, and linear models.

Zhi-Jun Liu is a Research Associate in the University Hygienic Laboratory at the University of Iowa. He received an M.S. in Statistics and a Ph.D. in Geography from the University of Iowa. His research interests include environmental modeling, spatial statistics, and geographic information systems.

George Hallberg is Associate Director of the University of Iowa Hygienic Laboratory and an Adjunct Professor in Geology and Civil and Environmental Engineering. He received his Ph.D. in Geology from the University of Iowa in 1975. Hallberg conducts research in environmental monitoring and public health surveillance.

# The Effect of Spatial Covariance Heterogeneity on Prediction Variance

J. Andrew Royle\* and Doug Nychka

---

**Abstract.-** Stationarity of a random field is an important assumption required for many spatial statistical analyses. A part of this assumption is homogeneity (translation invariance) of the correlation function. False assumption of homogeneity leads directly to misspecification of the spatial covariance and hence the usual kriging predictor is no longer BLUE. In this paper we examine the effect of homogeneous misspecification of a heterogeneous correlation function on prediction standard errors using two cases from a class of heterogeneous models that are spatially weighted mixtures of homogeneous random fields. It is found that misspecification leads to an average loss of efficiency in prediction by up to 20% for extreme cases, and that misspecification of the prediction variance generally leads to conservative prediction intervals.

---

## 1. INTRODUCTION

Kriging is a commonly used method for making spatial predictions. One of the primary reasons for the use of kriging over other methods of spatial prediction is that prediction standard errors are easily obtained for kriging and that the kriging predictor is BLUE. Unfortunately, the BLUE property holds only under the assumption that the spatial covariance is known, which is not likely to be the case in practice (i.e. there is generally a misspecification of the covariance function).

Kriging predictions are known to be fairly robust to misspecification of the spatial covariance structure (see Cressie and Zimmerman, 1992). Several authors have addressed the problem of stability of the kriging variance. In particular, Stein (1988, 1989) examined the effect of misspecification of the covariance structure on prediction variance. For a particular (fairly general) model, he derived a bound

---

\*Both Authors: Department of Statistics, North Carolina State University, Raleigh NC

on the ratio of the optimal prediction variance to the prediction variance for the predictor based on a misspecified covariance. More generally, he has shown that the misspecified prediction variance is asymptotically equivalent to the minimum prediction variance, assuming the true and misspecified covariance functions are *compatible* (see Stein and Handcock (1989) for a discussion of this condition).

Here we are interested in the departure of a random field from the assumption of translation invariance of the spatial correlation. Failure to detect departure from this assumption leads directly to misspecification of the covariance function, and we quantify the effects of this misspecification on prediction variance under a class of heterogeneous random field models.

## 2. STATIONARITY AND HETEROGENEITY

Let  $\mathbf{Z} = \{Z(x) : x \in \mathbb{R}^d\}$  be a random field in  $d$ -dimensions, from which we observe  $\mathbf{z} = \{z(x_i) : 1 \leq i \leq p\}$  where  $x$  represents a vector of coordinate locations in space (e.g.  $x = (x_1, x_2)$  in  $\mathbb{R}^2$ ).  $\mathbf{Z}$  is *second-order stationary* if

$$E(Z(x)) = \mu \quad \forall x \in \mathbb{R}^2$$

$$Var(Z(x)) = \sigma^2 \quad \forall x \in \mathbb{R}^2$$

and

$$Corr(Z(x), Z(x')) = \rho(x - x') \quad \forall x, x' \in \mathbb{R}^d$$

Thus under second-order stationarity we have that the mean and covariance *do not* depend on the location of the points in space (i.e. translation invariance of the first and second moments of  $\mathbf{Z}$ ). Concern over the validity of the stationarity assumption generally focuses on constancy of the mean and variance, however *homogeneity* (translation invariance) of the spatial correlation is also implied by stationarity. A heterogeneous random field then, is one in which the correlation function is *not* invariant to translation.

### 2.1 A Class of Heterogeneous Random Field Models

To examine the effect of heterogeneity on prediction variance it is necessary to specify a random field model with heterogeneous covariance structure. Here we consider special cases from the class of models that are spatially weighted mixtures of homogeneous random fields. The 2 component mixture is given by

$$Z(x) = a(x)U(x) + b(x)V(x). \tag{1}$$

Here the sampled field  $Z$ , is a spatially weighted sum of the homogeneous random fields  $U$  and  $V$  with covariance functions  $c_u$  and  $c_v$  and cross-covariance  $c_{uv}$ . The

nature of the spatial weights  $a(x)$  and  $b(x)$  determines the manner in which the covariance between two points depends on their relative locations. Note that the covariance between  $Z$  at  $x$  and  $x'$  under (1) is

$$\begin{aligned} Cov(Z(x), Z(x')) &= a(x)a(x')c_u(x-x') + b(x)b(x')c_v(x-x') \\ &\quad + (a(x)b(x') + a(x')b(x))c_{uv}(x-x') \end{aligned} \quad (2)$$

It is obvious that the covariance between two points depends explicitly on their locations  $x$  and  $x'$  via the weights  $a(x)$  and  $b(x)$ , and hence  $Z$  is a heterogeneous random field.

Simple cases of the mixture model are the heterogeneous variance model:  $Z(x) = a(x)U(x)$ , where  $Var(Z(x))$  is proportional to  $a(x)$  and the heterogeneous measurement error model:  $Z(x) = U(x) + b(x)V(x)$  where  $V(x)$  is measurement error with constant variance and  $Z(x)$  has measurement error (i.e. “nugget”) variance proportional to  $b(x)$ .

When the spatial weights are disjoint indicator functions (i.e. *either*  $U(x)$  or  $V(x)$  are observed at each  $x$ ) and  $c_{uv} = 0$ , then  $Z$  is homogeneous within subregions and  $Z$  has been called *pseudo-stationary*, or locally stationary. This model is equivalent to that implied by partitioning a region up into disjoint (homogeneous) subregions, a common solution to heterogeneity in classical geostatistics (Guttorp and Sampson, 1994). In this study, we consider the 2 subregion case further (*Model 1*, described below). A similar model is the *relative variogram model* of Cressie (1991) except that the latter assumes a proportional relationship between the covariance and mean within each subregion.

Taking the spatial weights to be overlapping (but not constant) produces a reasonable range of heterogeneous behavior, and this is the second situation that will be examined here (*Model 2*, described below). Moving-window kriging of Haas (1990) implies an underlying mixture model of this sort but with  $> 2$  components. Also related is the common (homogeneous) geostatistical “linear model of coregionalization” (Isaaks and Srivastava, 1989) which takes constant weight on each component of the mixture and hence the covariance is just a (weighted) sum of 2 or more covariance functions.

Let  $Z$  be defined over the region, denoted as  $\mathcal{X}$ , containing the  $6 \times 5$  lattice of sample points shown in Figure 1(a). Define subregions of  $\mathcal{X}$  as  $\mathcal{S} = \{x : x_2 < 3.5\}$  and  $\mathcal{N} = \{x : x_2 \geq 3.5\}$  and hence  $\mathcal{X} = \mathcal{S} \cup \mathcal{N}$ . The two subregions of  $\mathcal{X}$  are delineated by the broken line of Figure 1(a). The 2 heterogeneous models that we consider are defined as follows:

### **Model 1:**

Take  $U$  and  $V$  to be independent (i.e.  $c_{uv} = 0$ ) and to have isotropic exponential covariance functions, but with (possibly) different exponential parameters. So,  $c_u(\|x - x'\|) = e^{-\|x-x'\|\theta_u}$  and  $c_v(\|x - x'\|) = e^{-\|x-x'\|\theta_v}$ . Note that

$\theta_u = -\log(\rho_u(1))$  and  $\theta_v = -\log(\rho_v(1))$  so that we may describe the covariance functions of  $U$  and  $V$  in terms of the “nearest-neighbor” correlations  $\rho_u(1)$  and  $\rho_v(1)$ . These are the correlations between two points separated by distance 1 in Figure 1(a). The weighting scheme is defined as

$$a(x) = \begin{cases} 1 & \text{if } x_2 \in \mathcal{S} \\ 0 & \text{if } x_2 \in \mathcal{N} \end{cases}$$

and  $b(x) = 1 - a(x)$ . This case is relatively uninteresting, but serves as a “worst case” scenario, under which we might expect misspecification of the covariance structure to have the largest effect. The heterogeneous correlogram of  $Z$  on the  $6 \times 5$  lattice (which is equivalent to the covariogram in this case) is shown as the solid lines in Figure 1(b). One could generalize this situation to allow cross-correlation between  $U$  and  $V$ , and an example of such a covariance function is shown as the solid lines in Figure 1(c). Alternatively, we may induce a similar heterogeneous structure by overlapping the spatial weights, which gives us our second case.

### Model 2:

Taking  $U$  and  $V$  as in *Model 1*, build dependence among all  $Z(x)$  by superimposing the  $V$  field onto the  $U$  field (but only over part of the region, producing the heterogeneity). Here the weighting scheme takes  $a(x) = 1; \quad \forall x \in \mathcal{X}$  and

$$b(x) = \begin{cases} 1 & \text{if } x_2 \in \mathcal{S} \\ 0 & \text{if } x_2 \in \mathcal{N} \end{cases}$$

Thus, the covariance among points that both have  $x_2 \in \mathcal{S}$  contains the additive term  $c_v(\|x - x'\|)$ . The covariance among all other points is simply given by  $c_u(\|x - x'\|)$ . The correlogram of  $Z$  for this case is shown as the solid line in Figure 1(d). The heterogeneous models shown in Figures 1(c) and (d) are functionally very similar, as is evident. That is, they both build dependence among all  $Z(x)$ , but do this in a slightly different manner.

We will define the misspecified covariance function,  $c_m$ , to be the least-squares projection of the true covariance structure to a homogeneous exponential covariance (i.e. the least-squares fit to the model  $c_m(\|x - x'\|) = \beta_1 e^{-\|x - x'\|\beta_2}$ ). For *Model 1* and *Model 2*, the misspecified homogeneous covariance functions are shown as the broken lines in Figure 1(b) and (d), respectively.

## 3. STUDY DESIGN

Let  $Z$  be mixture of two homogeneous Gaussian random fields with  $E(U(x)) = E(V(x)) = 0$ . For this study, we choose a mixture of 2 homogeneous fields for convenience alone. The covariance between  $Z(x)$  and  $Z(x')$  is

$$Cov(Z(x), Z(x')) = c(x, x').$$

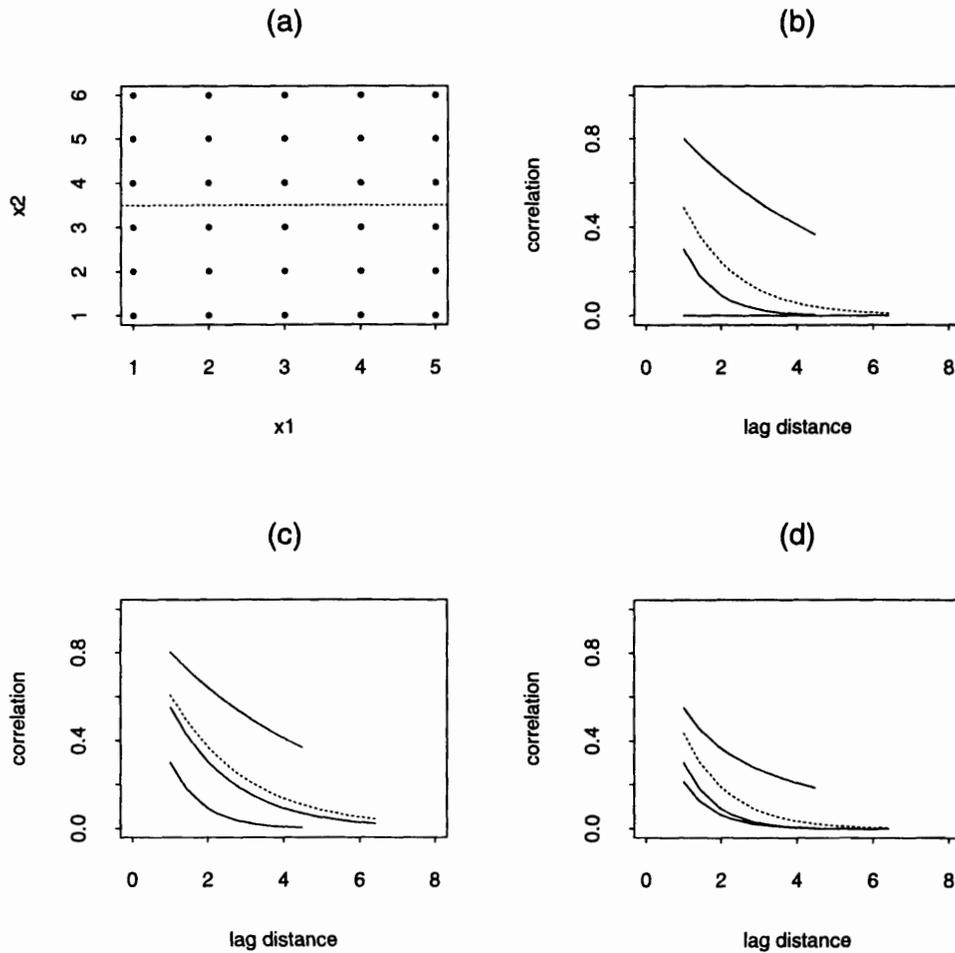


Figure 1: (a) - Location of 30 sample points used in studying the effect of heterogeneous correlation.  $\mathcal{X}$  is the region containing these points and the subregions  $\mathcal{S}$  and  $\mathcal{N}$  are delineated by the broken line. (b) - Heterogeneous correlation of 30 points under *Model 1* with  $c_{uv} = 0$ . (c) - Heterogeneous correlation of 30 points under *Model 1* with  $\rho_{uv}(1) = .55$ . (d) - Heterogeneous correlation of 30 points under *Model 2*. Broken line in all panels is the mis-specified homogeneous correlation function. For all cases,  $\rho_u(1) = .8$  and  $\rho_v(1) = .3$

For  $Z(\mathbf{x})$  corresponding to the  $p$  locations,  $\mathbf{x}$ , at which the random field is measured we have the  $p \times p$  matrix  $Var(Z(\mathbf{x})) = \mathbf{C}$  where  $\mathbf{C}_{ij} = c(x_i, x_j)$ . For any  $x_o \in \mathcal{X}$ , we have the  $p$ -vector  $Cov(Z(x_o), Z(\mathbf{x})) = \mathbf{c}_o$ . The  $p$  locations where  $Z$  is observed are given by the  $6 \times 5$  lattice shown in Figure 1(a).

For  $x_o \in \mathcal{X}$ , the kriging predictor of  $Z(x_o)$  is

$$\hat{Z}(x_o) = \lambda' \mathbf{z}$$

where the vector  $\lambda$  are the “kriging weights”, and are computed using the true covariance function defined by *Model 1* or *Model 2* or the misspecified covariance structure determined by  $c_m$ . These are the optimal or misspecified predictor, respectively, with kriging weights  $\lambda'$  and  $\lambda'_m$ .

The *prediction variance* is given by

$$Var(Z(x_o) - \hat{Z}(x_o)) = 1 - 2\lambda' \mathbf{c}_o + \lambda' \mathbf{C} \lambda$$

and the prediction standard error (PSE) is the square root of the prediction variance. The prediction variance for *either* the *optimal* or the *misspecified* predictor may be based on *either* the *true* heterogeneous covariance function,  $c$ , or the misspecified homogeneous covariance function,  $c_m$ . For the PSE associated with the prediction of  $Z(x_o)$ , 3 different estimates will be denoted as  $PSE_{opt}(x_o)$ ,  $PSE_r(x_o)$ , and  $MPSE(x_o)$ . The 3 estimators are defined as:

1.  $PSE_{opt}$  - The optimal estimator, using the true covariance function  $c$  in the predictor and prediction variance estimator.

$$PSE_{opt}^2(x) = 1 - 2\lambda' \mathbf{c}_o + \lambda' \mathbf{C} \lambda.$$

2.  $PSE_r$  - The correct estimator of PSE for the predictor based on the misspecified covariance,  $c_m$ .

$$PSE_r^2(x) = 1 - 2\lambda'_m \mathbf{c}_o + \lambda'_m \mathbf{C} \lambda_m.$$

3.  $MPSE$  - The *incorrect* estimator of PSE for the predictor based on the misspecified covariance,  $c_m$ .

$$MPSE^2(x) = 1 - 2\lambda'_m \mathbf{c}_{o,m} + \lambda'_m \mathbf{C}_m \lambda_m.$$

Note that the kriging weights, and covariance are indexed by  $m$  for the misspecified covariance function.

The estimator  $PSE_r$  (the subscript  $r$  for *reference*) serves as the “best possible” estimator given the incorrect predictor (i.e. using the predictor which is not optimal, but using the correct variance for the prediction error). Since  $PSE_{opt}$

is the optimal estimator of PSE, we know that  $PSE_r(x) > PSE_{opt}(x), \forall x \in \mathcal{X}$ . And, since  $MPSE$  is *not* the correct PSE estimator for the predictor based on the misspecified covariance, it may be that  $MPSE(x) < PSE_{opt}(x)$ .

We use these 3 estimators of PSE to compute prediction standard errors on a  $40 \times 40$  grid of points over the region containing the 30 sample points of Figure 1(a). The prediction standard errors are computed using only the 30 sample points, and we examine interesting summaries of the 3 PSE estimators. In particular, define the average *efficiency* of  $PSE_r$  to  $PSE_{opt}$  as

$$\text{Eff}(r, \text{opt}) = \frac{\sqrt{\int_{\mathcal{X}} PSE_r^2(x) dx}}{\sqrt{\int_{\mathcal{X}} PSE_{opt}^2(x) dx}}.$$

This is simply a measure of the relative precision of the misspecified predictor to the optimal predictor. Similarly, to quantify the worst precision over  $\mathcal{X}$  define the efficiency of the worst prediction as

$$\text{Eff}_{max}(r, \text{opt}) = \frac{\sqrt{\max_{x \in \mathcal{X}} PSE_r^2(x)}}{\sqrt{\max_{x \in \mathcal{X}} PSE_{opt}^2(x)}}.$$

These summaries provide an assessment of the predictor based on the misspecified covariance function. That is, under misspecification and assuming the correct prediction variance is used, they provide insight into the variability of the predictions relative to the optimal. However, supposing the heterogeneous covariance structure is misspecified, then the likely estimator of prediction standard error is  $MPSE$  and a useful measure of the cost of misspecification is the actual coverage probability for a 95% confidence interval for a prediction using  $MPSE$ , averaged over all predictions. For any  $x \in \mathcal{X}$ , the ratio  $MPSE(x)/PSE_r(x)$  is the relative inflation in the width of a confidence interval as a result of using the misspecified PSE estimator. The coverage probability of a 95% confidence interval for the prediction is  $CP(x) = Pr(Z(x) < MPSE(x)/PSE_r(x) \times 1.96) - Pr(Z(x) < -MPSE(x)/PSE_r(x) \times 1.96)$ . Hence, if  $MPSE(x)/PSE_r(x) = 1$ , the coverage probability is 0.95. The average *coverage* probability over  $\mathcal{X}$  is

$$ACP = \int_{\mathcal{X}} CP(x) dx.$$

Values of ACP *greater than* 0.95 indicate conservative confidence intervals (i.e. less precise predictions) and for ACP *less than* 0.95, predictions are too precise.

These summaries are computed for values of  $\rho_u(1) = .1, .2, \dots, .9$  and  $\rho_v(1) = .1, .2, \dots, .9$ , for each of the 2 models defined in Section 2.1 Recall that  $\rho_u(1)$  and  $\rho_v(1)$  are the nearest-neighbor correlations of the homogeneous  $U$  and  $V$  fields which comprise the heterogeneous mixture model. Thus, for each heterogeneous

case, there are 81 situations, ranging from very different dependence structure (e.g.  $\rho_u(1) = .1$  and  $\rho_v(1) = .9$ ), to identical (but independent) structure ( $\rho_u(1) = \rho_v(1)$ ,  $\rho_{uv}(1) = 0$ ). Efficiency and ACP will be presented as contour plots as a function of  $\rho_u(1)$  and  $\rho_v(1)$ .

#### 4. RESULTS AND CONCLUSIONS

Figure 2(a) indicates that the efficiency of prediction is effected very little by misspecification of the heterogeneous covariance structure under *Model 1*. That is, the mean prediction efficiency is near 1 except for high spatial dependence (nearest-neighbor correlations  $> 0.8$ ) in which case the prediction standard errors are only 20% larger for predictions based on the misspecified predictor. For the worst prediction under misspecification ( $\text{Eff}_{max}$ ), Figure 2(e) indicates inflation of PSE by 20 to 120% for levels of spatial correlation  $> 0.50$ . Since it is unlikely that one would use  $PSE_r$  as the estimator for prediction standard error if the heterogeneous covariance were misspecified, perhaps the more meaningful result is that concerning the average coverage probability using  $MPSE$ . The ACP is given in Figure 2(b), where it is seen that, using  $MPSE$  to estimate PSE, one achieves very conservative confidence intervals for most situations under *Model 1*. That is, the ACP for a 95% confidence interval is generally greater than 0.95, even approaching 1 for high spatial dependence.

For *Model 2*, Figure 2(c) indicates efficiency of the misspecified predictor near 1 and hence the misspecified spatial predictor is nearly optimal. In contrast to *Model 1*, Figure 2(f) indicates that for the worst predictions made under misspecification of *Model 2* the increase in maximum PSE for this case is only 20 to 40% for very high levels of spatial correlation (nearest-neighbor correlations  $> .80$ ) (compare with Figure 2(e)). Also in contrast to *Model 1*, the ACP using the misspecified estimator of PSE ( $MPSE$ ) under *Model 2* (Figure 2(d)) is generally *less than* 0.95 (i.e. liberal coverage). This occurs when nearest-neighbor correlations of the overlapping  $U$  and  $V$  are relatively very different (e.g.  $\rho_u(1) = .6$  and  $\rho_v(1) = .2$ ). For high and/or similar nearest-neighbor correlations, the ACP is  $\geq 0.95$  indicating conservative coverage of the confidence intervals for prediction.

By concentrating on a particular class of heterogeneous models for spatial correlation, defined to be a spatially weighted mixture of homogeneous random fields with exponential covariance functions, we have been able to quantify the effect of departure from homogeneity on spatial prediction uncertainty. One problem with studying the effect of heterogeneity, is that it is not clear how to describe the phenomenon in a manner so as to allow generalization of results, such as those presented here, to arbitrary covariance structure. Our choice of the mixture model is fortuitous since this is a heterogeneous analog to the class of models examined by Stein (1989) which are of the form

$$\text{Cov}(Z(s), Z(s')) = a_1 c_u(s - s') + b_1 c_v(s - s')$$

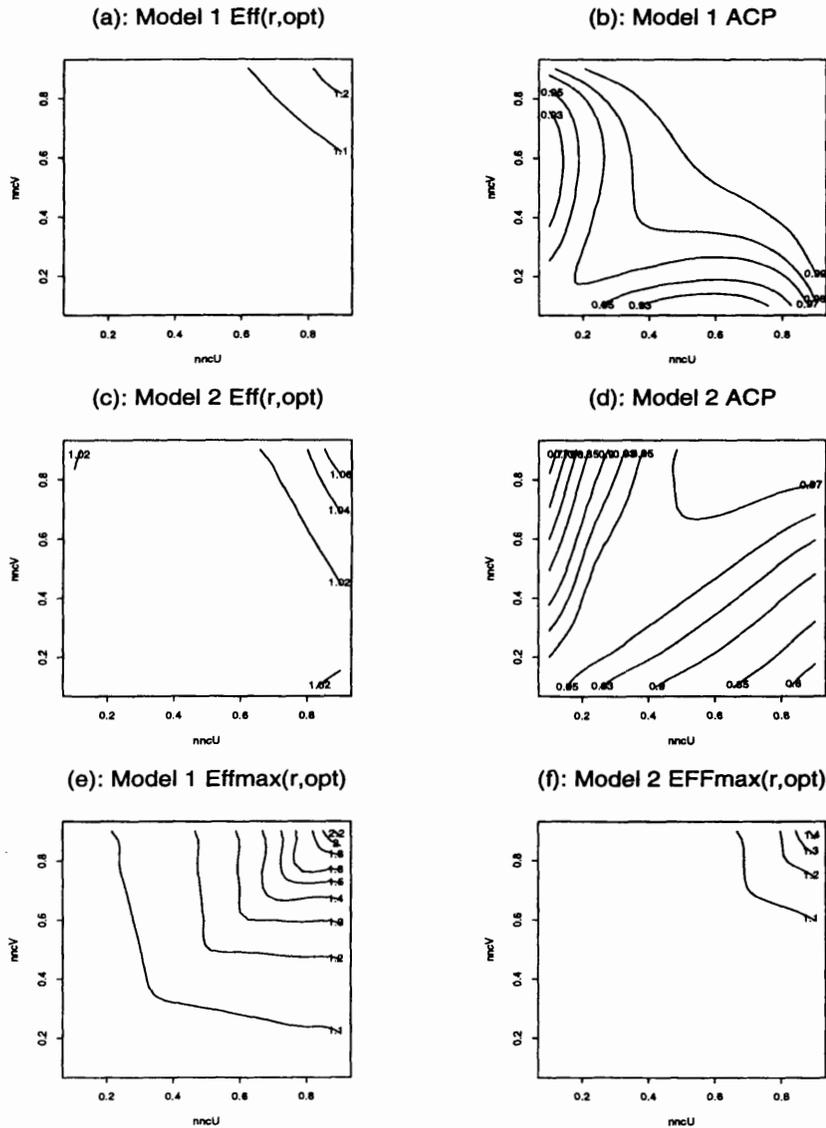


Figure 2: (a) Mean efficiency of the misspecified predictor under *Model 1*. (b) ACP under *Model 1* for 95% confidence intervals using the estimator of prediction standard error (*MPSE*) based on the misspecified (homogeneous) covariance. (c) Mean efficiency of the misspecified predictor under *Model 2*. (d) ACP using *MSPSE* under *Model 2*. (e) Efficiency of the least-precise prediction for the misspecified predictor under *Model 1*. (f) Efficiency of the least-precise prediction under *Model 2*. Axes are nearest-neighbor correlations of the  $U$  and  $V$  fields.

and were used by him to study the effect of misspecification (of stationary) covariance structure, and thus it is possible that results such as this may be extended to include heterogeneity of the form described here, but with an arbitrary covariance model (i.e. not restricted to the exponential class). This is an area of future work.

## REFERENCES

- Cressie, N.A.C. (1991). *Statistics for Spatial Data*. Wiley, New York, NY.
- Cressie, N.A.C. and D.L. Zimmerman (1992). On the Stability of the geostatistical method. *Math. Geol.* 24(1), 45-60.
- Guttorp, P. and P.D. Sampson (1994). Methods for estimating heterogeneous spatial covariance functions with environmental applications. in G.P. Patil and C.R. Rao, eds., *Handbook of Statistics* Volume 12, 661-689
- Haas, T.C. (1990). Lognormal and moving window methods of estimating acid deposition. *J. Amer. Stat. Assoc.* 85, 950-963.
- Isaaks, E.H. and R.M. Srivastava (1989). *An Introduction to Applied Geostatistics*. Oxford University Press.
- Stein, M.L. (1988). Asymptotically efficient prediction of a random field with misspecified covariance function. *Ann. Stat.* 16(1), 55-63.
- Stein, M.L. (1989). The loss of efficiency in kriging prediction caused by misspecifications of the covariance structure. in M. Armstrong, ed., *Geostatistics*. Kluwer Academic Publishers, 273-282.
- Stein, M.L. and M.S. Handcock (1989). Some asymptotic properties of kriging when the covariance function is misspecified. *Math. Geol.* 21, 171-190.

# Further Explorations of Relationships between Semi-Variogram and Spatial Autoregressive Models

Daniel A. Griffith, Larry J. Layne, and Philip G. Doyle<sup>2</sup>

---

**Abstract.**—This paper continues the forging of a common foundation for geo- and spatial statistics. To date partial conceptual correspondence has been established between the conditional autoregressive (CAR) and exponential semi-variogram models, the simultaneous autoregressive (SAR) and Bessel function semi-variogram models, and the moving average (MA) and linear semi-variogram models, when directional bias is not present. The exploratory numerical work summarized in this paper addresses the issue of whether or not these articulations are preserved when directional bias is present in the latent spatial autocorrelation, given symmetric spatial dependency. In doing so, impacts of edge effects are examined.

---

## INTRODUCTION

Generally speaking, spatial statistics is concerned with the statistical analysis of geo-referenced (or locationally tagged) data, and seeks to exploit the redundant information represented by spatial autocorrelation latent in such data. Two principal thrusts of spatial analysis have been (1) interpolation supporting visualization of surfaces, and (2) data analysis supporting statistical description and inference. One of the seminal visualization efforts was the famous SYMAP computer program, which sought to furnish low-resolution, coarse visualization of geographic surface representations (contour maps or perspective surfaces) based upon measurements made at a relatively small number of dispersed locations. Paralleling the development and dissemination of SYMAP was a body of work that focused on exploiting more localized spatial dependence latent in geo-referenced data, which has become known as the theory of regionalized variables, or geo-statistics. Matheron initiated this pioneering work in the 1960s, and Cressie (1991) recently published a state-of-the-art overview of geo-statistics. Meanwhile, spatial statistics research, especially characterized by Cliff and Ord's (e.g., 1981) work, focused attention on the need to employ spatially adjusted linear statistical models. Thus, while both geo- and spatial

---

<sup>1</sup>This research was supported by the National Science Foundation, research grant # SBR-9507855.

<sup>2</sup>Professor, Research Associate, and undergraduate research assistant (through an NSF REU supplement grant), respectively, Department of Geography, Syracuse University, 144 Eggers Hall, Syracuse, NY 13244-1090.

statistics support the visualization of geo-referenced data, spatial statistics goes beyond this task by enabling the drawing of sound statistical inferences from geo-referenced data.

While Davis and McCullagh (1975) began forging a common foundation for these two sub-disciplines, such a connection has continued to remain elusive. The principal objective of the research findings reported in this paper is to pursue this end. Its practical importance partly arises within the U.S. E.P.A. Environmental Monitoring and Assessment Program (EMAP) project, a major national endeavor in which statistics is being applied to environmental issues as well as monitoring and analysis undertakings, and partly in remediation work dealing with environmental contaminants. Concepts from both geo- and spatial statistics are necessary for successful engagement in these two scientific pursuits. Articulation of a common foundation for geo- and spatial statistics will better enable consistent statistical application and analysis of the massive amounts of geo-referenced environmental data becoming available. By doing so, geographic visualization and analysis will be able to more closely go hand in hand, and interpolated maps will be better able to be analyzed utilizing consistent spatial statistical model specifications.

### **Communalities Between Geo- and Spatial Statistics**

Historically, traditional statistical analyses have been concerned with data reduction and summarization. In contrast, time series analysis frequently leads to forecasting, whereas spatial series analysis often leads to interpolation, two data explosion operations. Geo- and spatial statistics similarly fixate on this interpolation issue through the "missing data" problem. Martin (1984) has outlined the exact maximum likelihood estimates of missing values for incomplete geographic data sets from a spatial statistics, autoregressive modeling perspective. The resulting equations, as demonstrated in Griffith (1993), are exactly those best linear unbiased estimates obtained with kriging (see Christensen, 1991), but in terms of the inverse covariance matrix rather than the covariance matrix itself. This link initially provided a conspicuous piece of evidence suggesting close relationships between semi-variogram and spatial autoregressive models, and furnished the motivation for Griffith and Csillag (1993) to numerically explore this relationship.

Algebraically speaking, then, estimation of missing data in spatial statistics (in order to avoid maps with holes) and kriging as an interpolator/predictor are exactly equivalent, which connects kriging to the E-M algorithm literature of statistics. Both numerical and graphical evidence exist to indicate that the spatial statistical conditional autoregressive (CAR) and geo-statistical exponential semi-variogram models are linked. Graphical evidence exists that couples the spatial statistical moving average (MA) and geo-statistical linear semi-variogram models. And, both numerical evidence and theoretical arguments signal a relationship between the spatial statistical simultaneous autoregressive (SAR) and the geo-statistical Bessel function (first order, second kind) semi-variogram models. The spectral density function of the SAR model may be written as:

$$P_{k,l} = \frac{\int_0^\pi \int_0^\pi \cos(k\theta_1) \cos(l\theta_2) [1 - 2\rho_p \cos(\theta_1) - 2\rho_q \cos(\theta_2)]^2 d\theta_1 d\theta_2}{\int_0^\pi \int_0^\pi [1 - 2\rho_p \cos(\theta_1) - 2\rho_q \cos(\theta_2)]^2 d\theta_1 d\theta_2}. \quad (1)$$

Christensen (1991) summarizes the status of this model by noting that Whittle (1954) has shown that the Bessel function can arise naturally in two-dimensional covariations, and the exponential semi-variogram model is a special case of it (Ripley 1981). The relevant functional form, where  $K_1$  is a modified Bessel function, is given by:

$$1 - \rho_{k,l} = c_0 + c_1 [1 - (\text{distance}/\text{range}) \times K_1(\text{distance}/\text{range})] . \quad (2)$$

When considering the appropriateness of this latter function as a descriptor of geographic covariation, numerical evidence (computed with Heuvelink's weighted least squares software) based on an isotropic dependency structure indicates that the two critical features leading to a Bessel function semi-variogram model specification appear to be intensity and degree of articulation of a geographic structure (see Table 1).

**Table 1. Bessel function semi-variogram fit of autoregressive models for geographic structures.**

Geographic articulation	Autoregressive model							
	conditional				simultaneous			
	$c_0$	$c_1$	range	SSD/SST	$c_0$	$c_1$	range	SSD/SST
square	0.350	0.617	2.56	0.007	0.005	0.994	8.27	0.000
hexagon	0.443	0.489	2.19	0.041	-----	-----	-----	-----
queen	0.442	0.520	2.88	0.004	0.011	0.988	10.1	0.000

Of note is that the exponential semi-variogram model (which actually is a Bessel function of order  $\frac{1}{2}$ ) furnishes a better description of the CAR cases than does the particular modified Bessel function estimated here. An SAR structure, which involves second-order dependencies, is better characterized by the popular Bessel function than is the CAR structure, which involves only first-order dependencies. Furthermore, relying upon an analogy with chess, the spatial linkage structure for the queen's set of neighbors (those areal units sharing any common boundary point) is better characterized by this Bessel function than is the rook's set of neighbors (only areal units sharing a common boundary of non-zero length). Moreover, as the spatial random field increases in scope (i.e., moving from a rook's to a queen's move dependency field), the SAR model more strongly relates to the Bessel function semi-variogram model. That is, the discrete setting more closely resembles the continuous setting. This tendency holds important implications for the U.S. E.P.A. EMAP database currently being constructed, which is being geo-referenced in terms of a hexagonal tessellation (comprised of approximately 12,600 hexagons) superimposed upon the continental U.S.

Numerical evidence presented in Table 1 was taken from exploratory computer experiments involving the following three general steps:

- Step 1: calculation of theoretical spatial correlations at various lags, for a regular square tessellation, based upon spatial autoregressive model specifications
- Step 2: fitting of selected geo-statistical semi-variogram models to the two-dimensional correlograms obtained in Step 1 using non-linear least squares
- Step 3: graphical superimposition of the estimated weighted least squares curves on the correlograms obtained in Step 1 using Heuvelink's computer software or customized SAS code.

The findings are restricted to translations from spatial statistical to geo-statistical model specifications, spotlighting the three overwhelmingly most popular spatial autoregressive models (i.e., the MA, CAR and SAR models).

### **Fortifying the Autoregressive-geo-statistical Linkages**

Research findings summarized here pertain to a further articulation of a common foundation for geo- and spatial statistics. One motivation for pursuing this work is the need for establishing a common umbrella for a growing body of seemingly disparate work (e.g., Rey, Getis and Bortman, 1994; Griffith, 1992; Little and Rubin, 1987). The extension outlined here builds on results reported by Griffith (1993) and Griffith and Csillag (1993). Two principal themes have been examined. One concerns non-stationary, multi-parameter spatial autoregressive models. Martin (1990) notes that the use of even bi-parametric spatial autoregressive inverse-covariance matrix specifications is extremely rare in applications, although well-established for very special cases. This avoidance of the two-parameter model in part is due to numerical difficulties affiliated with its estimation. The critical simplicity feature of a specification is that the pair of geographic weights matrices either are commutative or can exploit the matrix algebra notion of similarity. In contrast, specification and estimation of directional semi-variogram functions is quite common.

The second theme concerns boundary, or edge, effects. Griffith (1983) addresses the general problem of edge effects in quantitative geographical analyses, underscoring their seriousness. Rathbun (1994) provides an excellent example of the type of distortion overlooked edge effects can create and showed that properly handling them allowed better interpolation of salinity in Chesapeake Bay. In addition, results reported in Griffith and Csillag (1993) exhibit an affinity between the magnitude of prevailing spatial dependencies and the dominance of regional edge effects. This factor emerges because semi-variogram models directly deal with the covariance matrix for a set of observations, whereas spatial autoregressive models deal with the inverse of this matrix, and the procedure of matrix inversion is what amplifies and propagates edge effects. To date these effects have defied the formulation of general correction techniques. Insight into this edge effects phenomenon is gained here by holding the degree of spatial dependence constant, and changing the size of the region within which estimation is conducted (see Table 2).

**Table 2. Bessel function parameter estimates for queen's case, isotropic SAR specification,  $\rho = 0.2491$ .**

<b>Maximum lag</b>	<b><math>c_0</math></b>	<b><math>c_1</math></b>	<b>range</b>	<b>relative SS</b>
5	0.0093	0.906	9.35	0.000
10	0.0103	0.965	9.86	0.000
15	0.0109	0.978	10.00	0.000
25	0.0116	0.987	10.10	0.000
35	0.0119	0.987	10.10	0.000
5000	0.0200	0.987	10.50	0.000

Theoretically one would expect  $c_0 = 0$  and  $c_1 = 1$ . Notice that as the region used for

parameter estimation purposes increases in size, the fit of the Bessel function remains near-perfect, while the parameter estimates themselves change. The range of the spatial autocorrelation field increases to some upper limit and  $c_1$  approaches 1. Somewhat surprising is that  $c_0$  (the nugget effect) does not move toward zero, which may be indicative of a liability associated with using a non-linear model specification and/or weighted least squares estimation.

### Two-parameter Autoregressive Functions and Directional Semi-variogram Models

Theoretical correlograms can be constructed for the bi-parametric spatial autoregressive model matched with a regular square tessellation for the rook's set of neighbors spatial linkages since the closed form spectral density function is known. As noted previously, specification and estimation of directional semi-variogram functions is common. The theoretical two-dimensional correlograms generated by these equations do not need to be rotated, as their major and minor axes coincide with the corresponding horizontal and vertical axes.

Results of bi-parametric semi-variogram model fits are reported in Table 3. As has been found for the isotropic case, the exponential semi-variogram model relates almost perfectly to the spatial statistical CAR model case, and the Bessel function semi-variogram model relates almost perfectly to the spatial statistical SAR model case. For all cases explored here, the relative residual sums of squares essentially equals zero. This indicates that the linkages already established between the geo-statistical and spatial statistical models persist in the presence of anisotropy. Differences in the computed directional ranges are consistent with differences between the directional autoregressive parameters  $\rho_p$  and  $\rho_q$ . Furthermore, an inspection of the change in semi-variogram parameter estimates as the size of the region under study increases reveals the presence of edge effects, as it did in the isotropic situation. These edge effects become more pronounced as the two levels of spatial autocorrelation increase.

**Table 3. Parameter estimates and relative sums of squares (error sums of squares/corrected total sums of squares, or SSE/CTSS) for the CAR and SAR models for 4 different lattice sizes.**

		Autoregressive model					
		conditional			simultaneous		
lattice size	parameter	$\rho_1 = 0.05$ $\rho_2 = 0.25$	$\rho_1 = 0.175$ $\rho_2 = 0.320$	$\rho_1 = 0.050$ $\rho_2 = 0.445$	$\rho_1 = 0.05$ $\rho_2 = 0.25$	$\rho_1 = 0.175$ $\rho_2 = 0.320$	$\rho_1 = 0.050$ $\rho_2 = 0.445$
4 x 4	$C_0$	0.00000	0.13062	0.05397	0.02864	0.00922	0.00645
	$C_1$	1.00045	0.79249	0.89910	0.97118	0.93289	0.92101
	$b_1$	0.35341	1.52411	0.95340	0.35638	3.84913	1.93552
	$b_2$	0.76714	2.13413	3.08842	0.84330	5.25147	6.14246
	SSE/CTSS	0.00328	0.00338	0.00328	0.00225	0.00013	0.00059
6 x 6	$C_0$	0.00000	0.18844	0.09685	0.02906	0.01169	0.01215
	$C_1$	1.00020	0.76710	0.87825	0.97090	0.96549	0.96173
	$b_1$	0.35265	1.80853	1.07463	0.35662	4.00823	2.07320

	$b_2$	0.76649	2.51674	3.43566	0.84382	5.44961	6.42030
	SSE/CTSS	0.00309	0.00416	0.00554	0.00204	0.00006	0.00041
8 x 8	$C_0$	0.00000	0.22557	0.12686	0.02915	0.01335	0.01553
	$C_1$	1.00011	0.74630	0.85815	0.97083	0.97558	0.97252
	$b_1$	0.35239	2.00544	1.15218	0.35666	4.07235	2.12743
	$b_2$	0.76625	2.78110	3.66908	0.84392	5.52885	6.51884
	SSE/CTSS	0.00302	0.00509	0.00704	0.00197	0.00004	0.00032
10 x 10	$C_0$	0.00000	0.25132	0.14723	0.02919	0.01454	0.01771
	$C_1$	1.00007	0.72972	0.84279	0.97080	0.97932	0.97589
	$b_1$	0.35227	2.14681	1.20164	0.35668	4.10434	2.15386
	$b_2$	0.76615	2.97021	3.82763	0.84396	5.56817	6.56468
	SSE/CTSS	0.00298	0.00594	0.00804	0.00193	0.00003	0.00027

Theoretically the parameter estimates for  $C_0$  and  $C_1$  theoretically should equal 0 and 1, respectively. These are approximately the values estimated involving low levels of spatial autocorrelation, for both the exponential and Bessel function semi-variogram models. A tendency for  $C_1$  to move toward 1 is apparent for the intermediate and high levels of spatial autocorrelation in both model cases. Likewise a movement toward 0 is detectable for  $C_0$  in the exponential case but not in the Bessel function case.

#### **Empirical Case Studies: The Mercer-Hall Data and Pediatric Blood Lead Levels in Syracuse**

Two empirical data sets are investigated here. The first is the Mercer-Hall agricultural wheat-yield field plot data. These data were collected for 500 plots that formed a regular square 25-by-20 tessellation. Cressie (1991, p. 454) summarizes these data and their geographic landscape features. The second is pediatric blood lead levels in the City of Syracuse measured during the two-year period 4/1/92-3/31/94. This data set is comprised of 7,158 measurements (out of 15,277 cases for the County), some of which are replicates, spread across 3,841 locations. Twelve children had measures of zero, and for convenience were dropped from the analysis. Locational tags for these measures were obtained by address matching with ARC/INFO using TIGER files for Onondaga County, NY (with about an 80% success rate).

#### **The Mercer-Hall Data**

These data appear to be approximately normally distributed [ $P(\text{Wilk-Shapiro statistic} < 0.98041) = 0.0858$ ]. Besag (1974) fit a bi-parametric constant mean CAR model to these data, using a normalizing approximation proposed by Whittle (1954), and obtained the parameter estimates  $\hat{\rho}_p = 0.368$  and  $\hat{\rho}_q = 0.107$ . Exact maximum likelihood estimation (using IMSL subroutine UMINF) yields the estimates  $\hat{\rho}_p = 0.36445$  and  $\hat{\rho}_q = 0.11388$ , while using an extension of the Jacobian approximation outlined by Griffith and Sone (1995) implemented in SAS yields the estimates  $\hat{\rho}_p = 0.37634$  and  $\hat{\rho}_q = 0.12480$ . Both of these latter solutions began iterating from the OLS values of  $\hat{\rho}_p = 0.34045$  and  $\hat{\rho}_q = 0.13746$ . A mapping of the two-dimensional correlogram corroborated the need

for an anisotropic spatial statistical model specification, one that is symmetric and bi-parametric but not in need of an axis rotation. Restricting estimation of the exponential semi-variogram to 10 lags, and using weighted least squares and a model specification version following Haining (1990, p. 285), rendered the parameter estimates

$c_0$	$c_B$	range #1	range #2	SSD/SST
0.350	0.617	3.37411	0.74487	0.36600

The semi-variogram plot for these data, upon which the exponential model predictions are superimposed, is indicative of both a CAR and a bi-parametric model specification.

#### The Syracuse Pediatric Blood Lead Level Data

A log-transformed version of these data is closer to being approximately normally distributed [ $P(\text{Wilk-Shapiro statistic} < 0.9895) < 0.01$ ] than are the raw data themselves, and they also deviate far less from the homogeneous variance case. Because of the massive number of pairs of directional distances that can be computed for this dataset (i.e., 14,753,281) attention has been restricted here to distances less than or equal to 0.2 km. A two-dimensional correlogram for this range of distances indicates the need for a bi-parametric model specification involving an axis rotation. Results of such an exercise are contained in Table 4. The rotation angle,  $\theta$ , corroborates the need for an axis rotation in

**Table 4. Semi-variogram estimation for Syracuse blood lead levels.**

Model	$C_0$	$C_1$	range #1	range #2	relative SS	$\theta$
CAR	0.20396	0.58339	0.00192	0.01154	0.37078	0.04518 $\pi$
SAR	0.20832	0.57789	0.00070	0.04549	0.36326	0.03979 $\pi$

this empirical case. The range estimates corroborate the need for a bi-parametric model specification. Also, the SAR model furnishes a marginally better semi-variogram model fit.

### CONCLUSIONS

Additional numerical and empirical evidence is presented in this paper supporting the establishment of links between the CAR and exponential semi-variogram models, and the SAR and Bessel function semi-variogram models. This new evidence reveals that these links persist in the presence of anisotropy. Moreover, these articulations are preserved when directional bias is present in latent spatial autocorrelation, given directionally symmetric spatial dependency.

### REFERENCES

- Besag, J. 1974. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, 36: 192-225.
- Christensen, R. 1991. *Linear Models for Multivariate, Time Series, and Spatial Data*. NY: Springer-Verlag.

- Cliff, A., and J. Ord. 1981. *Spatial Processes*. London: Pion.
- Cressie, N. 1991. *Statistics for Spatial Data*. NY: Wiley.
- Davis, J., and M. McCullagh (eds.). 1975. *Display and Analysis of Spatial Data*. NY: Wiley.
- Griffith, D. 1983. The boundary value problem in spatial statistical analysis. *Journal of Regional Science*, 23:377-387.
- Griffith, D. 1992. Estimating missing values in spatial urban census data. *The Operational Geographer*, 10 (2): 23-26.
- Griffith, D. 1993. Advanced spatial statistics for analyzing and visualizing geo-referenced Data. *International Journal of Geographical Information Systems*, 7: 107-123.
- Griffith, D. and F. Csillag. 1993. Exploring relationships between semi-variogram and spatial autoregressive models. *Papers in Regional Science*, 72: 283-296.
- Griffith, D., and A. Sone. 1995 Trade-offs associated with normalizing constant computational simplifications for estimating spatial statistical models. *Journal of Statistical Computation and Simulation*, 51: 165-183.
- Haining, R. 1990. *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge: Cambridge University Press.
- Heuvelink, G. 1993. *Error Propagation in Quantitative Spatial Modelling*. Utrecht, The Netherlands: Faculteit Ruimtelijke Wetenschappen Universiteit Utrecht.
- Little, R., and D. Rubin. 1987. *Statistical Analysis with Missing Data*. NY: Wiley.
- Martin, R. 1984. Exact maximum likelihood for incomplete data from a correlated Gaussian process. *Communications in Statistics: Theory and Methods*, 13: 1275-1288.
- Martin, R. 1990. Spatial statistical processes in geographic modelling *in* *Spatial Statistics: Past, Present, and Future*, edited by D. Griffith, pp. 109-127. Ann Arbor, MI: IMAGe.
- Rathbun, S. 1994. Spatial modelling in irregularly shaped regions: kriging estuaries. paper presented at the annual joint statistical meetings, August 13-18, Toronto, Ontario, Canada.
- Rey, S., A. Getis, and A. Bortman. 1994. Spatial modeling approaches for the estimation of suppressed geo-referenced data. paper presented at the 41st North American meetings of the Regional Science Association International, Niagara Falls, Canada, November 17-20.
- Ripley, B. 1981. *Spatial Statistics*. NY: Wiley.
- Whittle, P. 1954. On stationary process in the plane. *Biometrika*, 41: 434-449.

### BIOGRAPHICAL SKETCH

Dr. Daniel A. Griffith is currently Chair, Department of Geography at Syracuse University. He holds a Ph.D. in Geography from the University of Toronto (1978), an M.S. in Statistics from The Pennsylvania State University (1985), an M.A. in Geography (1972) and a B.S. in Mathematics, both from Indiana University of Pennsylvania.

Larry J. Layne is a Ph.D. candidate at the College of Environmental Sciences and Forestry, S.U.N.Y. He holds an M.S. in Wildlife and Fisheries from South Dakota State University (1988) and a B.S. in Wildlife from Humboldt State University (1983).

Philip G. Doyle is a senior in the Department of Geography at Syracuse University and a G.I.S. Intern for Niagara Mohawk Power Corporation.

# A Gamma-function Model for D-variate Spectra and Cross-spectra for Large Scale Frequency Domain Simulation of Stationary Random Functions in $R^n$

Leon E. Borgman<sup>1</sup> and John W. Kern<sup>2</sup>

---

**Abstract.**--Spatially correlated random fields are used to model phenomena ranging in scale from description of the distribution of fibers in a sheet of paper to simulation of the forces of ocean waves along a beach. To make use of this framework with computationally efficient frequency domain methods, characterization of auto- and cross-spectra are required. Univariate and cross-spectral models based on the gamma probability density function are proposed for d-variate random vectors in n-dimensional space with geometric anisotropy. The corresponding class of spatial covariance functions is also derived. Modeling the second order structure between pairs of random variables is shown to simplify the usually tedious lag domain modeling of cross-covariance functions.

---

## INTRODUCTION

Spatially correlated random fields have been used to model physical phenomena ranging in scale from description of the distribution of fibers in a sheet of paper (Ahuja and Schachter, 1983), to simulation of the forces of ocean waves along a beach (Borgman et al. 1994). Prediction techniques usually known as kriging, (Krige 1951), Matheron (1971) and Journel and Huigbregts (1978) are used to produce best linear unbiased predictions at unsampled locations conditioned on known data. In contrast, conditional simulation methods (Journel and Isaaks 1984), (Borgman et al. 1984, 1993), are used to produce an ensemble of realizations of the random process where the correlation structure and natural variability of the original process are preserved and known data are interpolated.

Conditional simulations can be produced through space domain or frequency domain methods. However, for large scale regional simulations (100 by 100 grids or more), the space domain methods are slow and require large amounts of computer memory to execute. To overcome the limitations of space domain simulation

---

<sup>1</sup> Statistician, Department of Statistics and Geology/Geophysics, University of Wyoming, Laramie WY.

<sup>2</sup> Statistician, Western EcoSystems Technology Inc. Cheyenne WY.

techniques, the faster less memory-intensive frequency domain methods were developed, allowing simulation of larger spatially correlated random fields at much greater speeds.

To obtain the spectral density models required for frequency domain techniques, covariance functions developed for use in space domain applications such as the spherical model, hole effect model or members of the Matern class (Matern 1960), have typically been Fourier transformed to frequency domain. In general, these spectral density models have little intuitive connection with space domain processes and in some cases demonstrate pathologies in practice.

With the exception of Goulard and Voltz (1992), there has been little development of cross covariance/spectral density models for simulation of multivariate correlated random fields. Since few environmental or ecological problems are univariate in nature, this lack of available multivariate models has limited application of conditional simulation techniques to these areas.

We propose a class of univariate and multivariate spectral/covariance models suitable for simulation of regional scale processes. The class of models is flexible and easily interpreted using properties of the individual univariate processes.

### UNIVARIATE SPECTRAL MODEL WITH INTUITIVE PARAMETERS

The proposed models will be developed for second order stationary random functions (Cressie 1991, page 53). Given any  $d$  dimensional covariance function  $C(\mathbf{h})$  corresponding to a second order stationary random function such that  $\int |C(\mathbf{h})| d\mathbf{h} < \infty$ , there exists a  $d$ -dimensional spectral density function given by the Fourier transform (Bochner 1955)

$$\begin{aligned} S(\mathbf{f}) &= \mathcal{F}^{-1}(C(\mathbf{h})) = \int_{\mathbf{R}} C(\mathbf{h}) e^{-i2\pi \mathbf{f}^T \mathbf{h}} d\mathbf{h} \\ C(\mathbf{h}) &= \mathcal{F}^{+}(S(\mathbf{f})) = \int_{\mathbf{R}} S(\mathbf{f}) e^{+i2\pi \mathbf{f}^T \mathbf{h}} d\mathbf{f}. \end{aligned} \tag{1}$$

A spectral density function  $S(\mathbf{f})$  is isotropic if  $S(\mathbf{f})=G(|\mathbf{f}|)$  for some scalar function  $G$ , and radial if there exists some  $d$  dimensional linear transformation  $A$ , such that  $S(\mathbf{f})=G(|A\mathbf{f}|)$ . In other words a space with radial spectral density function can be deformed through linear transformation to an isotropic space.

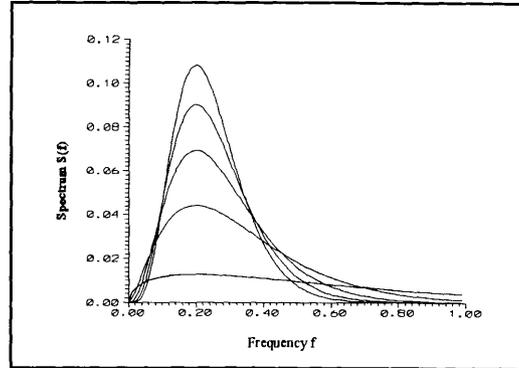
We propose an anisotropic radial spectral model with an ellipsoidal base, and radial function based on the gamma probability density

$$G_n(\rho) = K_n \rho^{\alpha-1} e^{-\frac{\rho}{\beta}}, \tag{2}$$

where  $\rho$  represents frequency scaled in ellipsoidal coordinates. A subset of the family of gamma radial functions is given in figure 1.

### Intuitive Motivation of Parameters

At large scales, many physical processes exhibit nearly-periodic behavior (i.e. the human eye observes repetition although the pattern also has a random component). Examples include topography of mountainous terrain, vertical height of ocean waves, recurrence of vegetative patches and porosity of fractured media. Simulations of these processes at large scales should include these “pseudo-periodicities”. The mode frequency of the gamma spectral density function controls the periodicity of a simulation. If the spectrum is very peaked, ( $\alpha \gg 1$ ), most of the variation will be concentrated in oscillations at a single frequency. The simulation will be nearly perfectly periodic like an egg carton. In contrast, if the spectrum is very flat ( $\alpha \approx 1$ ), the corresponding simulated data will contain oscillations at nearly all frequencies (white noise). Anisotropies controlled by the ellipsoidal base of a radial spectrum as in figure (2) cause “ridges” in simulated data. Few physical processes are isotropic (one rarely sees round mountains) and hence this feature is essential to realistic simulations. The axes of the ellipsoidal base are easily chosen or estimated from data. Finally the scaling constant  $K_n$  may be chosen to give simulated data the proper variance. Using the gamma spectral density model in frequency domain, these properties are easily controlled to produce conditional or unconditional simulations. The same properties are implicit in space domain methods although not directly controlled by the practitioner.



**Figure 1. Family of radial gamma spectral models with mode frequency 0.2 and  $\alpha$  ranging from 1.5 to 5.5 and  $\beta = 1/(\alpha-1)$ . These models all have dominant wavelength 5 units with varying amounts of high frequency noise.**

### COVARIANCE FUNCTIONS CORRESPONDING TO RADIAL SPECTRA

Let  $S(\mathbf{f})$  be a radial spectral density function with radial function  $G$ , corresponding to a second order stationary random function with  $\mathbf{f} \in \mathbb{R}^n$  ( $n$  dimensional Euclidean space), and ellipsoidal base given by  $\mathbf{f}^T B \mathbf{f} = 1$ , where  $B = V L V^T$ , is the eigenvalue/eigenvector decomposition of  $B$ . Making the substitutions  $\boldsymbol{\phi} = L^{1/2} V^T \mathbf{f}$ ,  $|J| = |B|^{-1/2}$  and  $\mathbf{h} = V L^{1/2} \boldsymbol{\tau}$ , in equation (1) gives

$$C(\mathbf{h}) = \int_{\Omega} G(\sqrt{\Phi^T \Phi}) e^{i2\pi \Phi^T \tau} |B|^{-1/2} d\Phi. \quad (3)$$

The n-dimensional integral in equation (3), can be reduced to a one dimensional Hankel transform through introduction of the hyper-spherical coordinate transformation (Miller 1964 ). Computing n-1 integrals gives

$$C(\mathbf{h}) = \frac{(2\pi)^{\frac{n}{2}}}{R^{(n-1)/2}} \int_0^{\infty} \rho^{\frac{n-1}{2}} G(\rho) J_{(n-2)/2}(R\rho) \sqrt{R\rho} d\rho, \quad (4)$$

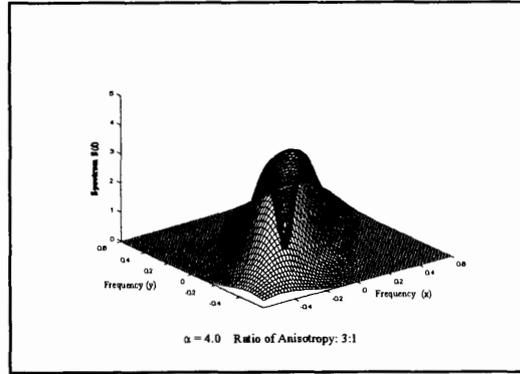
where  $R = \sqrt{\mathbf{h}^T B^{-1} \mathbf{h}}$ , is a function of  $\mathbf{h}$ , and  $J_v$  is the Bessel function of order  $v$ . This is the Hankel transform of order  $(n-2)/2$  which is tabled for many common functions by Ditkin and Prudnikov (1965 p. 432). Note that  $C$  is a function of  $|L^{-1/2} V^T \mathbf{h}|$  and hence must be radial with an ellipsoidal base. Further, the covariance function is given by the Hankel transform of the radial function . These results for second order stationary random functions can be combined with the converse from Taheri (1980) and Hagen (1982).

In summary, we have that for  $S(\mathbf{f})$  and  $C(\mathbf{h})$ , the spectral density/covariance pair with radial functions  $G$  and  $g$  respectively;  $S(\mathbf{f})$  is radial with ellipsoidal base  $\mathbf{f}^T B \mathbf{f} = 1$  if and only if  $C(\mathbf{h})$  is radial with ellipsoidal base  $\mathbf{h}^T B^{-1} \mathbf{h} = 1$ , and

$$C(\mathbf{h}) = \frac{2\pi}{R^{(n-2)/2}} \int_0^{\infty} \rho^{(n-1)/2} G(\rho) J_{(n-2)/2}(2\pi\rho R) \sqrt{2\pi\rho R} d\rho \quad (5)$$

if and only if

$$S(\mathbf{f}) = \frac{2\pi}{\rho^{(n-2)/2}} \int_0^{\infty} R^{(n-1)/2} g(R) J_{(n-2)/2}(2\pi\rho R) \sqrt{2\pi\rho R} d\rho. \quad (6)$$



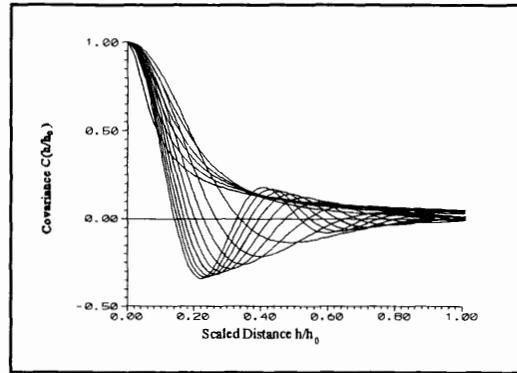
**Figure 2. Two dimensional radial gamma spectral density function with 3 to 1 anisotropy.**

## The Gamma Spectral Model for n Dimensional Space

These results can be used to derive the gamma and other spectral density covariance pairs in n dimensions. Using equation (5) and the Hankel transform of the gamma radial function (Ditkin and Prudnikov 1965, p.439, eq. 11.38), the n dimensional covariance function corresponding to the gamma spectral model is given by

$$C_n(\mathbf{h}) = K_n \frac{2\pi\beta^{\frac{n+2\alpha}{2}}}{\sqrt{|B|R^{n-2}}} \cdot \frac{\Gamma(n+\alpha+1)}{(1+(2\pi\beta R)^2)^{(n+2\alpha)/4}} \cdot P_{\frac{n-2}{2}}^{\frac{n-2}{2}} \left( \frac{1}{(1+(2\pi\beta R)^2)^{1/2}} \right), \quad (7)$$

where  $P_{\nu}^{\mu}(z)$ , is the Legendre Spherical function (Stegun p. 332, eq. 8.1.2) and the scaling constant  $K_n = \sqrt{|B|} \sigma^2 \cdot (2\beta^{\alpha+n-1} \Gamma(\alpha+n-1) \Gamma(n/2) \pi^{n/2})^{-1}$ , is obtained by forcing the integral of the spectrum over the full frequency space to be  $\sigma^2$ . A subset of the class of radial functions is plotted in figure (3). It should also be noted that the class of covariance functions is quite flexible containing members with shapes ranging from that of the Gaussian to the hole effect model. Widely varying model types can be chosen through parameter estimation alone as opposed to fitting several models from separate classes followed by ad hoc procedures for choosing the best class of models.



**Figure 3. Family of covariance functions corresponding to the gamma spectral model with  $\alpha$  ranging from 0.01 to 0.61 in increments of 0.2 and from 1 to 46 in increments of 5.**

### MULTI-VARIATE CROSS-SPECTRA IN $\mathbb{R}^n$

To simulate multivariate (vector) random processes with frequency domain methods, models of cross spectra are required. In general these models are obtained through modeling the cross covariance in lag domain. This is sometimes unsatisfactory in that the cross covariance is typically not symmetric about zero and does not have any a-priori expected form. However, with some simplifying assumptions, a reasonable cross spectral model can be obtained which is intuitive in nature. In particular, it is assumed that the coherence and phase are approximately constant in the range of frequencies with significant power in the univariate spectra.

Let  $\mathbf{x}$  be an n dimensional vector in  $\mathbb{R}^n$ , and  $V_k(\mathbf{x})$ ,  $k=1,2,3\dots d$  be a set of d

intercorrelated second order stationary random functions. The cross covariance function between the  $ij^{\text{th}}$  pair of random functions is

$$C_{ij}(\mathbf{h}) = E\{[V_i(\mathbf{x}) - \mu_i][V_j(\mathbf{x} + \mathbf{h}) - \mu_j]\} \quad (8)$$

For  $i = j$ , this is the auto covariance function. The cross spectrum for any pair is

$$S_{ij}(\mathbf{f}) = \int_{\Omega} C_{ij}(\mathbf{h}) e^{-i2\pi\mathbf{f}^T\mathbf{h}} d\mathbf{h} = c_{ij}(\mathbf{f}) - iq_{ij}(\mathbf{f}) \quad (9)$$

where  $c_{ij}$  and  $q_{ij}$  are the co- and quad-spectra. In general,  $C_{ij}(\mathbf{h})$  may not be symmetric about the origin, and hence  $S_{ij}(\mathbf{f})$  will be complex valued. However, since  $C_{ij}(\mathbf{h})$  is real valued, the co-spectrum is even and the quad-spectrum is odd (Borgman, 1993). The coherence and phase are defined in terms of the co- and quad-spectra as

$$\gamma_{ij}^2(\mathbf{f}) = \frac{c_{ij}^2(\mathbf{f}) + q_{ij}^2(\mathbf{f})}{S_{ii}(\mathbf{f})S_{jj}(\mathbf{f})}; \quad \phi_{ij}(\mathbf{f}) = \arctan\left(\frac{q_{ij}(\mathbf{f})}{c_{ij}(\mathbf{f})}\right). \quad (10)$$

These relations can be looked at as a polar transformation with angle  $\phi_{ij}$  and radius  $r = \gamma_{ij}\sqrt{S_{ii}S_{jj}}$ . Solving for  $c_{ij}$  and  $q_{ij}$  gives

$$c_{ij} = \gamma_{ij}\sqrt{S_{ii}S_{jj}} \cos\phi_{ij} \quad \text{and} \quad q_{ij} = \gamma_{ij}\sqrt{S_{ii}S_{jj}} \sin\phi_{ij} \quad (11)$$

where the argument  $\mathbf{f}$  has been suppressed (although assumed to be present) for notational convenience. Note that if  $S_{kk}$  is zero, then the co and quad- spectra are also zero. Therefore values for  $\gamma_{ij}$  and  $\phi_{ij}$  are required only for those frequencies where the individual univariate spectra are nonzero. The proposed model is

$$S_{ij}(\mathbf{f}) = \begin{cases} \gamma_{ij}\sqrt{S_{ii}(\mathbf{f})S_{jj}(\mathbf{f})} e^{-i\phi_{ij}}, & \text{for } F_L \leq \mathbf{f}^T B \mathbf{f} \leq F_U \quad f_1 \geq 0 \\ \gamma_{ij}\sqrt{S_{ii}(\mathbf{f})S_{jj}(\mathbf{f})} e^{i\phi_{ij}}, & \text{for } F_L \leq \mathbf{f}^T B \mathbf{f} \leq F_U \quad f_1 < 0 \\ 0 & \text{elsewhere} \end{cases} \quad (12)$$

where the phase and coherence are assumed constant in the range of frequencies where the univariate spectra are non-zero, and where  $F_L$  and  $F_U$  are the lower and upper bounds on that region. The model is defined piecewise to allow the quad-spectrum to be odd. The sets on the right side of equation (12) are denoted  $\Omega^+$  and  $\Omega^-$  respectively. These sets could be generalized to any half space.

To apply the model, estimates of  $S_{ii}$ ,  $S_{jj}$ ,  $\gamma_{ij}$ , and  $\phi_{ij}$  are required. The marginal spectra may be estimated and modeled independently using appropriate spectral

models. The gamma is proposed here due to its flexibility in shape and ease of application, although other spectral models may be used.

Given that each of the  $d$  marginal spectra have been estimated or reasonably chosen, the coherence and phase remain to be estimated. By expressing the cross-covariance functions in terms of the inverse Fourier transform of the cross-spectra, the phase and coherence parameters (assumed constant) may be estimated in lag domain. The Fourier transform of the cross spectra may be expressed as a sum of sine and cosine integrals  $C_{ij}(\mathbf{h}) = \gamma_{ij} \cos(\phi_{ij}) I_c(\mathbf{h}) + \gamma_{ij} \sin(\phi_{ij}) I_s(\mathbf{h})$ , where  $I_c$  and  $I_s$  are given by

$$I_c \equiv \int_{\Omega} \sqrt{S_{ii} S_{jj}} \cos(2\pi \mathbf{f}^T \mathbf{h}) d\mathbf{f},$$

and

$$I_s = \int_{\Omega^-} \sqrt{S_{ii} S_{jj}} \sin(2\pi \mathbf{f}^T \mathbf{h}) d\mathbf{f} - \int_{\Omega^+} \sqrt{S_{ii} S_{jj}} \sin(2\pi \mathbf{f}^T \mathbf{h}) d\mathbf{f}. \quad (13)$$

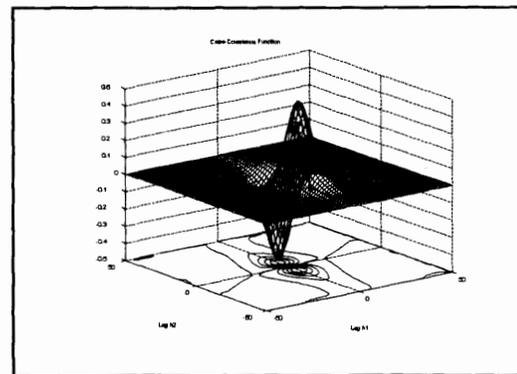
If  $C_{ij}(\mathbf{h})$  has been estimated possibly with some form of bin averaging or nonparametric smoothing at a set of  $N$  lags  $\{\mathbf{h}_k^*, k=1,2,3,\dots,N\}$ , estimates of the phase and coherence may be obtained from the  $N$  linear equations in  $\gamma_{ij} \cos\phi_{ij}$ , and  $\gamma_{ij} \sin\phi_{ij}$ ,

$$\hat{C}_{ij}(\mathbf{h}_k^*) = \gamma_{ij} \cos\phi_{ij} I_c(\mathbf{h}_k^*) + \gamma_{ij} \sin\phi_{ij} I_s(\mathbf{h}_k^*), = g_{ij1} I_c(\mathbf{h}_k^*) + g_{ij2} I_s(\mathbf{h}_k^*). \quad (14)$$

These equations can be solved by least squares to obtain estimates of  $g_{ij1}$  and  $g_{ij2}$ . The estimated phase and coherence are given by

$$\hat{\gamma}_{ij} = \sqrt{\hat{g}_{ij,1}^2 + \hat{g}_{ij,2}^2} \quad \hat{\phi}_{ij} = \arctan\left(\frac{\hat{g}_{ij,2}}{\hat{g}_{ij,1}}\right). \quad (15)$$

An example of a cross correlation function corresponding to the proposed gamma model is plotted in figure (4). Each of the marginal processes have univariate gamma spectra with ellipsoidal bases. The maximum correlation is located  $270^\circ$  counter clockwise from the first lag domain axis  $h_1$ , which corresponds to a constant phase shift of  $270^\circ$ . The minimum correlation is negative and located  $180^\circ$  from the maximum correlation. This type of cross



**Figure 4. Cross covariance function for 2 correlated anisotropic random functions with radial gamma spectral models, coherence 1.0 and phase angle  $270^\circ$ .**

correlation function could be used to simulate the relation between rainfall and surface runoff where a directional lag would be anticipated in the spatial correlations.

## DISCUSSION

Frequency domain conditional simulation techniques allow the user to directly control pseudo-periodicities anisotropy and high frequency noise in simulated data sets. The univariate gamma spectral density model allows the user to easily control these characteristics by directly identifying parameters with physical properties. The gamma model has been applied by and Peacock and Kern (1995) to model hydraulic conductivity in the Powder river basin in Wyoming at regional scales.

The proposed multivariate models provide a framework for estimating cross spectra or equivalently cross-correlation functions in which complicated lag domain functions are simplified to a pair of band limited functions of frequency which are typically zero over most of the frequency domain. Further research into this area should include investigation of non-constant functional forms for the phase and coherence.

Frequency domain methods should be of particular interest to geographers and environmental scientists handling large data sets in GIS systems. Large scale simulations at computational speeds obtained with frequency domain methods are not feasible with space domain procedures.

## ACKNOWLEDGMENTS

We are grateful for funding of this work which came from a cooperative agreement between the state of Wyoming Department of Environmental Quality, Wyoming State Engineers Office, the United States Office of Surface Mines, the United States Bureau of Land Management and the University of Wyoming.

## REFERENCES

- Bochner, S. 1955. *Harmonic Analysis and the Theory of Probability*. University of California Press, Berkeley and Los Angeles, CA.
- Borgman, L.E. and R.C. Faucette, 1993. Basic mathematics and statistical theory for finite Fourier coefficients of Gaussian vector random functions. In: *Computational Stochastic Mechanics*, Chap. 1, Computational Mechanics Publ., London.
- Borgman, L.E., C.D. Miller, S.R. Signorini and R.C. Faucette, 1994. Stochastic interpolation as a means to estimate oceanic fields. *Atmosphere-Ocean*, 32(2) 1994, p. 395-419.
- Borgman L.E., Taheri, M., and R. Hagan, 1984. Three-dimensional, frequency-domain simulations of geological variables. In *Geostatistics for Natural Resources Characterization, Part 1*, G. Verly, M. David, A.G. Journel, and A. Marechal, eds. Reidel, Dordrecht, p. 517-541.

- Cressie, N.A.C. 1991. *Statistics for Spatial Data*, John Wiley And Sons, Inc. New York.
- Ditkin, V.A. and A.P. Prudnikov, 1965. *Integral Transforms and Operational Calculus*. Translated by D.E. Brown and edited by I.N. Sneddon. Pergamon Press, New York.
- Hagan, R. 1982. Application of spectral theory and analysis in mining geostatistics and statistical linear wave theory. PhD. Dissertation, Department of statistics, University of Wyoming, Laramie WY.
- Journel, A. and C.J. Huijbregts, 1978. *Mining Geostatistics*, Academic Press, New York.
- Journel, A. and E.H. Isaaks. 1984. Conditional simulation: applications to a saskatchewan uranium deposit. *Mathematical Geology*, 16:685-718.
- Krige, D.G. 1951. A statistical approach to some mine valuation and allied problems at the Witwaterstrand. Unpublished masters thesis, University of Witwaterstrand.
- Matern, B. 1986. *Spatial Variation*, In Lecture Notes in Statistics No 36, Edited by: D. Brillinger, S. Fienberg, J. Gani, J. Hartigan and K. Krickeberg, Springer Verlag, NewYork.
- Matheron, G. 1971, The theory of regionalized variables and it's application. Re. No. 5, Les Cashiers de centre de morphologie mathématique de fontainbleau, C.G. Fontainbleau.
- Peacock, K. and J. W. Kern. 1995. Assessment of groundwater impacts related to the proposed Lighthouse coal bed methane project. United States Department of Interior, Bureau of Land Management, Casper District. Casper WY.
- Stegun, I.A. 1972. Legendre functions, In: *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*. Edited by M. Abramowitz and I.A. Stegun. p 331-354.
- Taheri, S. M., 1980. Data Retrieval and Multidimensional Simulation of Mineral Resources. PhD Dissertation, University of Wyoming, Department of Statistics, Laramie Wyoming.

### **BIOGRAPHICAL SKETCH**

Leon E. Borgman is a professor of statistics and geology and geophysics at the University of Wyoming, Laramie, Wyoming. He has specialized in application of stationary random function theory to a wide range of environmental science, hydrgeologic and oceanographic/civil engineering problems.

John W. Kern is a consulting statistician with Western Ecosystems Technology Inc., Cheyenne, Wyoming. John has primarily specialized in application of stationary random function theory to environmental, hydrogeological and ecological problems.

# Computing the Area Affected by Phosphorus Runoff in an Everglades Wetland Using Bayesian and Universal Kriging

Song S. Qian

---

**Abstract.** --Phosphorus-enriched agriculture runoff is believed to be the leading cause of ecosystem changes in an Everglades wetland. In order to study the effects of the added nutrients to the wetland ecosystem, it is necessary to estimate the acreage of the affected region. In this study, Bayesian and universal kriging are used to analyze the data collected by Reddy, *et al* (1991). The background level of the STP concentration is used as an indicator of whether the region is affected or not through an indicator function. The expected value of the affected area is calculated using the results from Bayesian and universal kriging. The results show that universal kriging is sensitive to the specification of covariance function. Universal and Bayesian kriging yield comparable results when the covariance functions are specified in the like manner.

---

## INTRODUCTION

The Water Conservation Area-2A (WCA2A) is part of the Everglades wetlands in south Florida (Figure 1). It receives the agricultural runoff from the Everglades Agriculture Area. The phosphorus enriched agriculture runoff caused some significant changes in this phosphorus limited wetland ecosystem. The most obvious change was in the change of the dominant vegetation species from sawgrass to cattail near the inlet of the runoff. The significant increase of phosphorus level in the water of WCA2A is considered a major threat to the Everglades National Park. To protect the Park, the South Florida Water Management District proposed to use constructed wetlands, a man-made marsh like buffer area, to remove excess phosphorus before the water enters WCA2A.

One important design parameter of a constructed wetland is the unit mass loading rate. The unit area loading rate is usually calculated by measuring the influent mass loading rate (in mass of phosphorus per unit time) and dividing it by the area of the receiving wetland. Because WCA2A covers such a large region, not all parts of the region are effective in removing phosphorus. (In other words, not all parts are affected.) If the entire WCA2A were taken as affected and used to calculate the mass loading rate, the phosphorus assimilating capacity of the wetland is underestimated.

The data used in this paper are from Reddy *et al.* (1991), who sampled soils at 74 stations on seven north -- south transects, spaced at 2 mile intervals (Figure 2). Richardson (personal communication, 1993) believes that the STP content should be fairly stable if there were no agricultural runoff problem, and the background level of the STP content is

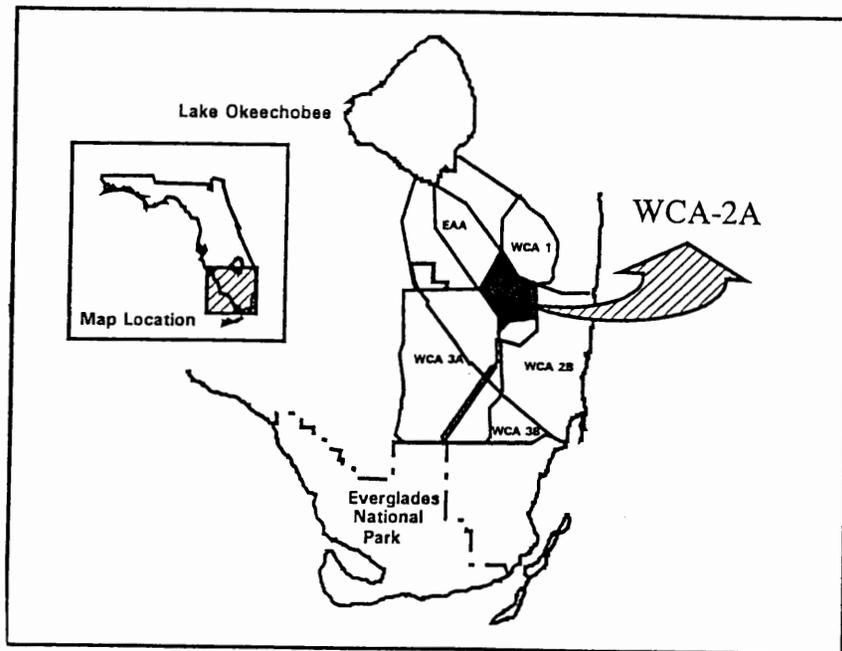


Figure 1. Map of south Florida and location of the study region

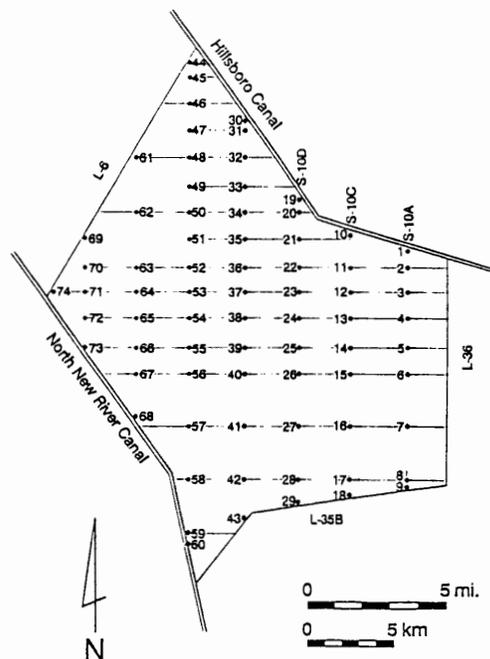


Figure 2. Locations of sampling sites in WCA2A

about 500 µg of phosphorus per gram of soil (µg/g) in the top 20 cm. Universal kriging (Cressie, 1991) and Bayesian kriging (Handcock and Stein, 1993) are used to model the phosphorus concentration over the region. When the concentrations are larger than the background level, it is believed that the corresponding region is affected by the agricultural runoff.

## METHODS

### Data analysis

The mean STP content in the top 20 cm layer of the Reddy samples were used. The STP values are log-transformed to stabilize the variance. The spatial coordinates were converted from latitude and longitude to the Universal Transverse Mercator (UTM) grid system

Universal kriging (UK) can be interpreted as a Gaussian random field model:

$$Z = X\beta + \eta \quad (1)$$

where:  $Z$  is the log transformed STP content,  $X$  is the known design matrix containing the longitude and latitude, and the distance to the nearest pollution source,  $\beta$  is a vector of unknown parameters, and  $\eta$  is the error term which is assumed to be normally distributed with mean 0 and covariance  $\Sigma$ . The covariance function is  $cov\{Z(s_1), Z(s_2)\} = \alpha K_\theta(s_1, s_2)$ , for any pair of spatial coordinate points  $(s_1, s_2)$  of interest, where  $\alpha > 0$  is a scale parameter, and  $\theta$  is a vector of structural parameters which specify the shape of the covariance function. In the case of kriging, we observe  $\{Z(s_1), Z(s_2), \dots, Z(s_n)\}' = Z$  and the prediction of  $Z(s_0)$  is the objective, where  $s_0$  is a location with no observed data. The UK predictor is the best linear unbiased predictor (Cressie, 1991), of the form  $\hat{Z}_\theta(s_0) = \lambda(\theta)'Z$ . See Ripley (1981) for details.

The covariance matrix  $\alpha K_\theta$  is usually unknown, and is often estimated using the empirical semi-variogram. Once the semi-variogram model is chosen, the covariance function is taken as known, and the uncertainty in the selected model is ignored. To include this uncertainty in the analysis, Handcock and Stein (1993) introduced Bayesian kriging (BK) using the Matern class of covariance functions, with a parameter  $\theta = (\theta_1, \theta_2)$ .

The posterior predictive distribution for  $Z(s_0)$  is derived by adapting a noninformative prior:

$$pr(\alpha, \beta, \theta) \propto pr(\theta)/\alpha \quad (2)$$

Handcock and Stein (1993) showed that the conditional posterior distribution of  $Z(s_0)$  is:

$$Z(s_0)|\theta, Z \sim t_{n-q} \left( \hat{Z}_\theta(s_0), \frac{n}{n-q} \hat{\alpha}(\theta) V_0(\theta) \right) \quad (3)$$

where  $q$  is the number of parameters in  $\beta$ .

The marginal posterior distribution of  $\theta$  is:

$$pr(\theta|Z) \propto pr(\theta) |K_\theta|^{-1/2} |X'K_\theta^{-1}X|^{-1/2} \hat{\alpha}(\theta)^{-(n-q)/2} \quad (4)$$

The Bayesian predictive distribution for  $Z(s_0)$  is:

$$pr(Z(s_0)|Z) \propto \int_{\Theta} pr(Z(s_0)|\theta, Z) \times pr(\theta|Z)d\theta \quad (5)$$

where:  $\hat{Z}_\theta(s_0)$  is the UK point predictor,  $\hat{\alpha}(\theta) = \frac{1}{n-q}(Z - X\hat{\beta})'K_\theta^{-1}(Z - X\hat{\beta})$ ,  $\alpha V_0(\theta)$  is the UK prediction error variance, and  $t_{n-q}$  is a student- $t$  distribution with  $(n - q)$  degrees of freedom. See Handcock and Stein (1993) for details.

The prior distribution of the parameters of the covariance function was chosen as:

$$pr(\theta) = pr(\theta_1) \times pr(\theta_2) = \frac{1}{(1 + \theta_1)^2(1 + \theta_2)^2} \quad (6)$$

This prior distribution represents the belief that large values of the parameters are less likely than small ones.

### Area estimation

The expected value of the area can be estimated by:

$$E(A_p) = \int [\int I(Z(X) > t)dX] \times pr(Z(X) | Z)dZ = \int pr(Z(X) > t | Z)dX \quad (7)$$

where  $I(\cdot)$  is an indicator function,  $X$  is the coordinate vector, and  $t$  is the background level.

When UK is used, the integrand in the RHS of equation (7) can be evaluated by using the CDF of a normal density. When BK is used:

$$E(A_p) = \int_{X \in R^2} \int_{\Theta} pr(Z(X) > t | \theta, Z) \times pr(\theta | Z)d\theta dX \quad (8)$$

The uncertainty of the estimated area using BK is evaluated using a Monte Carlo simulation method. The procedure is based on the fact that given  $\theta$ , the area is  $A_p = \int I(Z(X) > t|Z, \theta)dX$ . It can be approximated by sampling  $Z(X)$  and evaluating the integrand, and the mean of these values is approximately equal to the proportion of the affected area over the entire WCA2A. A Monte Carlo simulation algorithm was used to sample  $\theta$  and then  $Z(X)$ .

## RESULTS

The residuals of the model (1) are found to be intrinsically stationary and isotropic through visual inspection of the semi-variograms in the north-south and east-west directions. For UK, four different covariance functions are used. One is the Matern covariance function using the maximum posterior estimate of  $\theta$  (UK-M). The other are: exponential (UK-exp), spherical (UK-sph), and Gaussian (UK-Gau). Figure 3 shows the log-transformed data.

### Covariance function

The covariance function used in BK is presented in Figure 4, the log transformed posterior as a function of the parameters ( $\theta_1$  and  $\theta_2$ ). The posterior

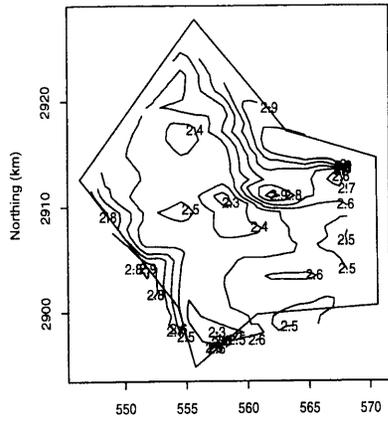


Figure 3. Log-transformed data

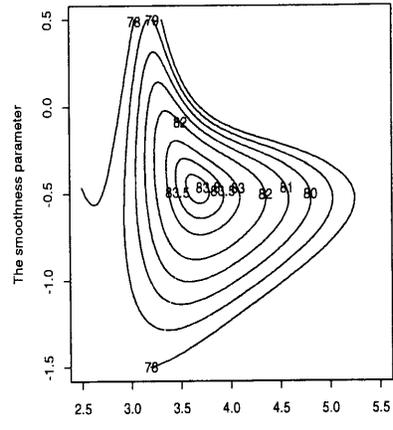


Figure 4. Posterior of the covariance function parameters

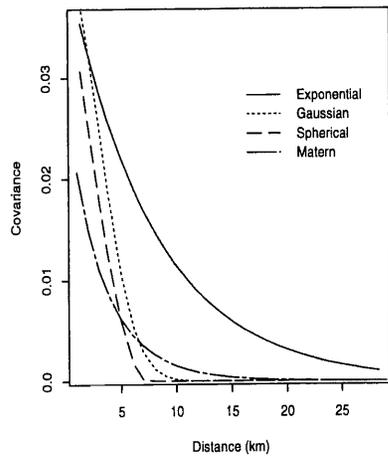


Figure 5. Covariance functions

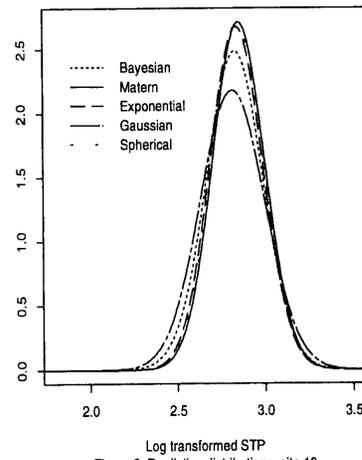


Figure 6. Predictive distributions, site 10

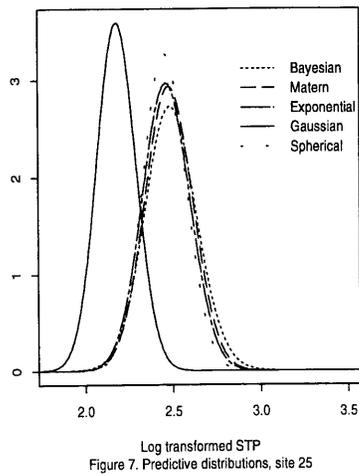


Figure 7. Predictive distributions, site 25

mode is at (3.6515, -0.4697), or, the Maximum Posterior Estimates (MPE) of the parameters are (4482, 0.3391). The estimated  $\alpha$  given the MPE of  $\theta$  is 0.032363.

The covariance functions in UK are estimated through fitting the semi-variogram model. Figure 5 compares the covariance functions computed from different semi-variogram models. In Figure 5, the line labeled "Matern" is the Matern covariance function using the MPE of  $\theta$ .

### Predictive distributions

The predictive distributions of the log-transformed STP are estimated for the 10th sampling site, which has the largest STP content, and the 25th sampling site, which has a below average STP content. The density functions are shown in Figures 6 and 7. Table 1 compares the results from BK and UK using the three semi-variogram models.

**Table 1. Predictive distributions**

Model	Sampling Site	Predicted Mean	Prob. > 500 $\mu\text{g/g}$	Measured STP
BK	25	2.48754	0.076779	2.398536
UK-M	25	2.478177	0.052068	2.398536
UK-exp	25	2.462711	0.030408	2.398536
UK-Gau	25	2.179858	0.000001	2.398536
UK-sph	25	2.453698	0.022335	2.398536
BK	10	2.826391	0.786981	3.100039
UK-M	10	2.850576	0.848553	3.100039
UK-exp	10	2.83751	0.823649	3.100039
UK-Gau	10	2.814715	0.736275	3.100039
UK-sph	10	2.843363	0.835841	3.100039

Comparing the five predictive distributions for the 10th sampling site, we note that the predictive distributions from UK-exp, UK-sph, and UK-Mat are nearly identical. BK yields a similar mean value but slightly larger estimation variance. The predictive distribution from UK-Gau has the largest variance. For the 25th sampling site, we see comparable predictive distributions from UK-exp, UK-sph, and UK-M. BK yields a similar mean and slightly larger variance. UK-Gau produces a much different predictive distribution from the other four. The mode is significantly less than the modes from the other four predictive distributions. The Gaussian covariance function is equivalent to the Matern covariance function with  $\theta_2 \rightarrow \infty$ . This means that UK-Gau describes a function which is infinitely differentiable. This representation may not be realistic.

### Computation of the affected area

Equation (6) was used for UK and Equation (7) was used for BK. The results are listed in Table 2.

In this study, we did not try to locate the boundary of the affected region. However, the shape of the affected region may be stable and we plot the contour

lines of the probability of STP larger than 500  $\mu\text{g/g}$  to indicate the possible shape. Figure 8 is plotted using BK, and Figures 9 to 12 show the surfaces from UK. The surface from the UK-Gau is mostly flat but has sudden jumps. The reason for this unnaturally shaped surface is that the Gaussian covariance function describes "super smooth" surfaces, in this case, super smooth for the residuals. A model with very smoothed residual surface must have a jumpy mean surface.

**Table 2. Expected area**

Model	affected area (km <sup>2</sup> )	% of the total area
BK	97.186	22.394
UK-M	95.775	22.069
UK-sph	89.403	20.601
UK-exp	99.615	22.954
UK-Gau	76.588	17.648

From Table 2, we note that the estimated area using BK is very close to the estimated area using UK-M and UK-exp. The estimated area is sensitive to the specification of the covariance function when UK is used. It is reasonable to state BK is more appropriate in this study, since the covariance function is unknown and the uncertainty in estimating the covariance function is considered.

The uncertainty of the estimated area is evaluated only for BK. Figure 13 presents the histogram of the percentages of the affected area from the Monte Carlo simulation.

## DISCUSSION

In conclusion, it is seen that one significant difference is an increased computational intensity of BK. However, the results indicate that UK is sensitive to the specification of the covariance function. The difference between two different covariance functions may be significant even though the semi-variogram models of the two are fitted to the data equally well. This sensitivity to the covariance function justifies the use of BK.

There are several studies that estimated the phosphorus affected area of WCA2A. Craft and Richardson (1993) use the same data set as presented in this paper. In their study, the boundary of the affected region was delineated by using both the STP content and the phosphorus accumulation rate. The delineated boundary is plotted onto a USGS topographic map, and the area is thus measured. Their estimate of the area is 115 km<sup>2</sup>. Only the top 10 cm layer of the Reddy data was used and 600  $\mu\text{g/g}$  is the background level. The STP contour lines were produced using the ordinary kriging algorithm from a commercial graphic software (Surfer<sup>®</sup>). The semi-variogram model used is not reported.

Using the STP content as an indicator of the influence of the agricultural runoff is a more appropriate way of estimating the affected area. A spatial statistical

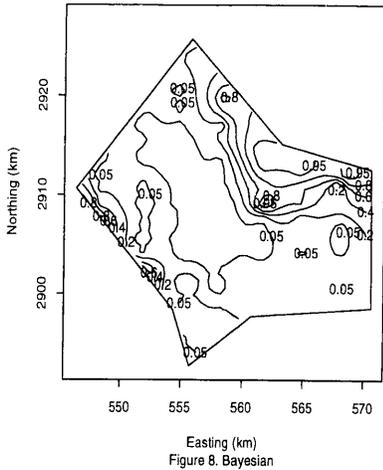


Figure 8. Bayesian

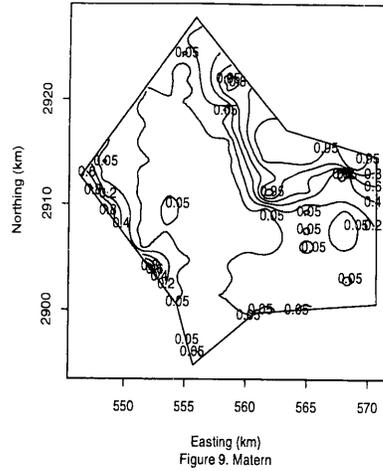


Figure 9. Matern

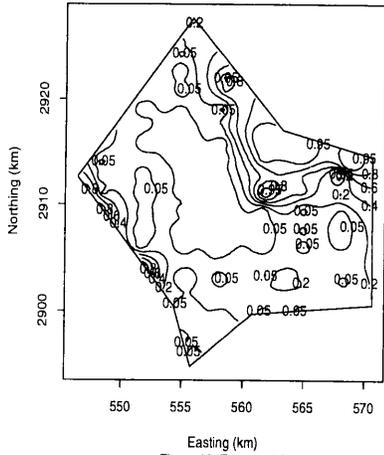


Figure 10. Exponential

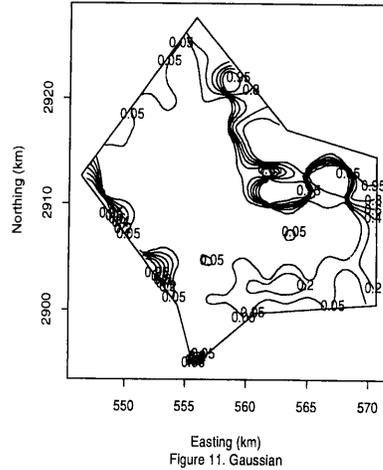


Figure 11. Gaussian

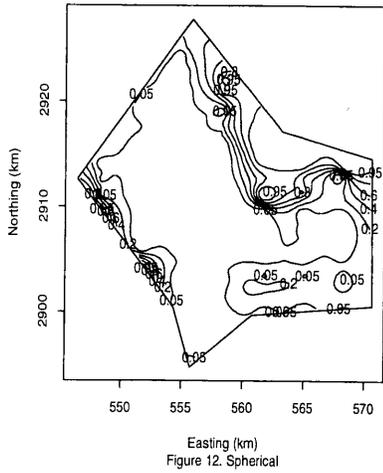


Figure 12. Spherical

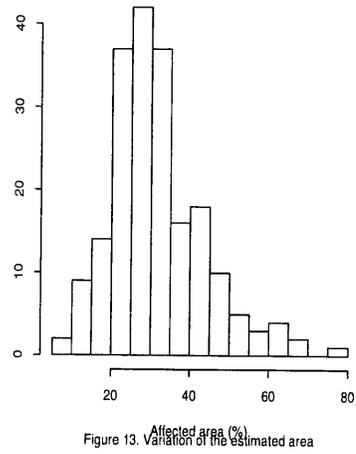


Figure 13. Variation of the estimated area

approach accounts for the uncertainty in the sample data. Considering the high level of uncertainty involved in the estimating process (Figure 13), the estimated area using BK should be considered comparable to the result in Craft and Richardson (1993). The discrepancy between the estimates may be caused by two factors, without considering the error introduced by the process of measuring the area from a map. They are their use of ordinary kriging and may be a different covariance function, and the different background level. There is no reason to believe that the background level is a constant. The field samples from the region that is unaffected by the agriculture runoff show a significant variation (Craft and Richardson, 1993). In a current study by the author, the background level is represented by a probability distribution.

### ACKNOWLEDGMENT

Drs. Richard Smith and Peter Müller motivated the author to pursue this project. Dr. Marcia Gumpertz's comments and suggestions significantly improved this article. The author appreciate the encouragement and support from Drs. C.J. Richardson and K.H. Reckhow.

### REFERENCES

- Craft, C.B. and C.J. Richardson. 1993. Peat accretion and phosphorus accumulation along a eutrophication gradient in the northern Everglades. *Biogeochemistry*, 22, 133-156.
- Cressie, N.A.C., 1991. *Statistics for Spatial Data*, John Wiley & Sons, Inc. New York.
- Handcock, M.S., and M.L. Stein. 1993. A Bayesian analysis of kriging, *Technometrics*, 35(4), 403-410.
- Howard-Williams, C. 1985. Cycling and retention of nitrogen and phosphorus in wetlands: a theoretical and applied perspective, *Freshwater Biology*, 15, 391-431.
- Reddy, K.R., W.F. DeBusk, Y. Wang, R. DeLaune, and M. Koch, 1991. *Physico-chemical properties of soils in the Water Conservation Area 2 of the Everglades*. Final report submitted to South Florida Water Management District, West Palm Beach, FL 33416.
- Ripley, B.D. 1981. *Spatial Statistics*, John Wiley & Sons, Inc. New York.

### BIOGRAPHICAL SKETCH

**Song S. Qian** is a postdoctoral research associate at the Duke Wetland Center of the Nicholas School of the Environment of Duke University. He received his Ph.D. in environmental sciences and MS degree in statistics from Duke University. He has a Master's degree in environmental systems engineering from Nanjing University, China.

# Mapping Synecological Coordinates: A Spatial Analysis of Environmental Indices in a Forested Landscape

Margaret R. Holdaway<sup>1</sup> and Gary J. Brand<sup>2</sup>

---

**Abstract.**—The Method of Synecological Coordinates was used to compute environmental indices (synecological coordinates) for moisture and nutrients based on plant species presence. Plots were sampled in a 65.6-ha area in St. Louis County, Minnesota, dominated by quaking aspen. A four-level, nested sampling design quantified spatial variability over a wide range of distances. Semivariations at the four distances suggested a 45-m equidistant grid of plots would be appropriate for mapping synecological coordinates. Variograms of moisture and nutrients coordinates showed ranges of 350 m or greater, indicating that spatial dependence continues well beyond the limits normally found with most soil studies. Nutrients coordinate showed greater spatial dependence than moisture coordinate, which showed more random variation. The nutrients coordinate map proved useful in defining areas of differing quaking aspen site index. Thus, the efficient mapping of site synecological coordinates has helped capture the pattern of environmental variation.

---

## INTRODUCTION

Our ability to appropriately manage a forest community can be improved if we can accurately map environmental conditions. We propose a general method that maps environmental indices (plant-based indicators). The intent of this paper is to investigate whether environmental indices can be accurately mapped, and if so, whether the maps identify areas of contrasting site productivity.

Environmental indices are produced by the Method of Synecological Coordinates (MSC) (Bakuzis 1959, Bakuzis and Kurmis 1978, Brand 1985, and Gutiérrez-Espeleta 1991). MSC uses plant presence as an indicator of the environment. It is based on the notion that floristic composition of a site indicates the nature of the environment there. MSC is an easily applied approach that computes semi-quantitative values for four environmental factors—moisture (M), nutrients (N), heat (H), and light (L). M and N represent the edaphic conditions and H and L represent the climatic conditions at the site. Each plant species has its own set of synecological coordinate values on a relative scale of from 1 to 5, representing on average, the species “need” for that factor. An arithmetic mean of the coordinates for those species occurring together is called the site synecological

---

<sup>1</sup> *Mathematical Statistician, North Central Forest Experiment Stn, St. Paul, MN.*

<sup>2</sup> *Research Forester, North Central Forest Experiment Stn, St. Paul, MN.*

coordinate for that factor. For this paper we will concentrate on edaphic environmental conditions as represented by M and N.

Synecological coordinates, which use the integrative nature of plants and prior knowledge of their environmental "requirements," may provide a practical means to determine the spatial variability and distribution of environmental factors. Spatial variability is analyzed using geostatistical techniques. Geostatistical theory is based on the simple notion that measurements made closer together in space are more alike than those made farther apart. The spatial variability or structure can be modeled via a mathematical function known as the variogram. Once the variogram is modeled, kriging uses information in the variogram to spatially interpolate unknown data points.

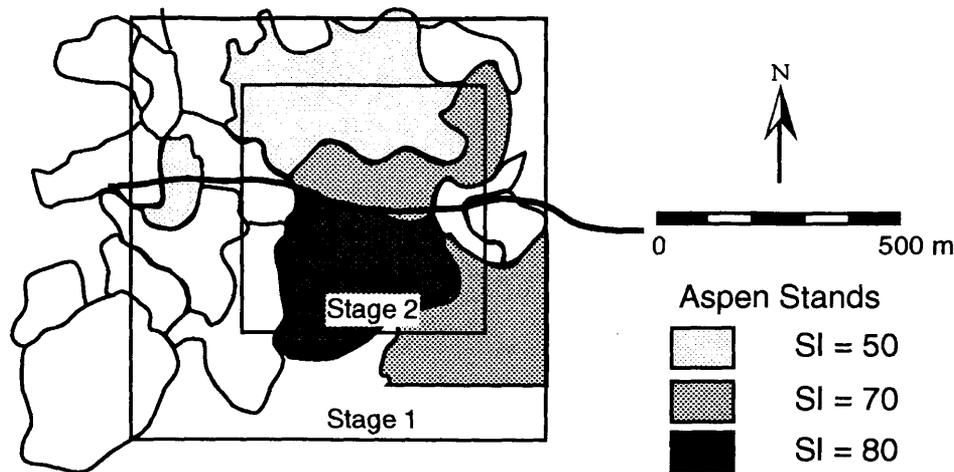
The spatial variation in soil properties has been extensively analyzed with geostatistical techniques (Oliver and Webster 1991, Burgess and Webster 1980, and Webster and Burgess 1983). Because soil is such an important determinant of the environmental conditions of a site, describing the spatial distribution of soil properties should spatially describe important aspects of the environment. However, the chemical and physical composition of soils is complex and extremely variable over short distances (Lechowicz and Bell 1991, Jackson and Caldwell 1993, Oliver and Webster 1991). Therefore, synecological coordinates may provide a more effective means to integrate soil variability and complexity into maps of edaphic conditions.

The choice of sampling distance to use in any study depends on the range over which the phenomena are spatially correlated. Complex multi-step designs are especially suited for exploring the full spatial structure when there is no previous knowledge of the spatial nature of the data. Designs containing several sampling scales, such as a systematic-cluster design (Fortin et al. 1989) or a nested design (Oliver and Webster 1986) are ideally suited for such investigations. Sampling is performed in two stages. In the initial stage, a complex sampling scheme is carried out over a wide range of distances to ascertain the range of the process. This information then helps establish the sampling grid in the second stage, where points are sampled on an equal-spaced grid at a scale well within the spatial range.

We are investigating synecological coordinates and geostatistics as methods to represent the local edaphic environment in an aspen-dominated forest. By representing the edaphic environment, the medium of forest growth, we expect to make inferences about overstory site productivity. Our purpose here is: (1) to analyze the spatial structure and scale of spatial variability of site synecological coordinates for moisture and nutrients using geostatistical techniques, (2) to assess the accuracy of site coordinates mapped by kriging, and (3) to examine what the mapped site coordinates indicate about site productivity.

## METHODS

The stage 1 study site (larger square in figure 1) covers 65.6 ha in St. Louis County, 30 km north of Duluth, Minnesota. A timber inventory (based on limited field sampling) delineated stand boundaries and measured site index (SI), the traditional measure of site quality in forest management. Quaking aspen is the primary overstory species. The topography is gently rolling with elevations ranging from about 1440 feet to slightly more than 1490 feet with no obvious directional trends in the topography.



**Figure 1.—Aspen-dominated landscape in St. Louis County, Minnesota. Heavy line shows road, light lines delineate stand boundaries, and squares delineate study areas. Shading shows the site index class (SI) of the aspen stands.**

A nested sampling design was selected to provide variability data for a wide range of distances. In stage 1 of the sampling design, we determined the variability at coarse (135 m), moderate (45m), fine (15 m) and very fine (5m) distances. The sampling distances needed to be simultaneously small enough to estimate the nugget variance (unexplained or random variance that occurs at very small sampling distances) and large enough to identify the spatial range.

In stage 1, the primary or coarsest grid of points (level 1) had separation distances of 135 m. The axes of the sampling grid were oriented in a north-south, east-west direction with the starting grid point randomly located. A geometric progression of decreasing distances between sample points was used. Therefore, the distances for the second-level points were 45 m. Two second-level points were located 45 m to the south and west of each first-level point (see figure 2a). Each primary and secondary point included third- and fourth-level points (figure 2b). The two third-level points were located 15 m to the south and west of every primary and secondary point; the two fourth-level points were located 5 m to the north and east of every primary and secondary point.

Collecting plant lists at each point in a complete four-level sampling design would have been too intensive for the time and money available. Therefore, alternating primary rows and columns were dropped from the study for second-level points (figure 2a) and the number of third- and fourth-level points were reduced by randomly selecting half of the level 1 and level 2 points for sampling at levels 3 and 4. At each point, plant species occurring within a square plot of 5x5 m were recorded and synecological coordinates were computed using a data recorder and program (Nimerfro and Brand 1993). Species synecological coordinates developed for Minnesota forests were used in the computations (Bakuzis and Kurmis 1978).

Obvious outliers were removed from the data at each stage. A plot was removed if its value for either factor was judged an outlier by the box plot. The stage 1 semivariance based on sampling at four lag distances (5, 15, 45, and 135

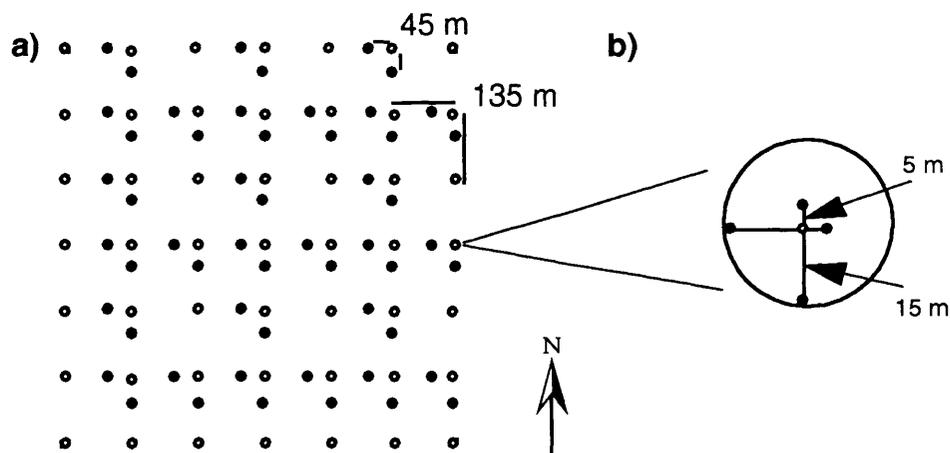


Figure 2.—Final sampling design: (a) coarse and moderate distances, (b) fine and very fine distances.

m) was plotted against distance. Ranges were estimated, and an appropriate sampling distance was chosen from among the four lag distances for stage 2 plot spacing.

In stage 2, the experimental unidirectional variograms were calculated. Cross validation was performed to check the validity of the final models. Differences between kriged estimated and observed values were analyzed using these cross validation statistics: mean error, median absolute error, and MSE.

### VARIOGRAM MODELING AND CROSS VALIDATION

Site synecological coordinates were computed from 254 plots at stage 1 grid points. At an additional 13 plots, plant lists could not be collected due to environmental conditions detrimental to terrestrial vegetation (road, permanent standing water, etc.) or site coordinates were outliers. The semivariances at the four lag distances for stage 1 data (black dots on figure 3) show increasing spatial variability at 135 m for N. However, this distance may be too large when kriging M, so we chose 45 m as the grid sampling distance for the stage 2 analysis.

Stage 2 plots (smaller square in figure 1) were located on a 11x11 grid centered within the stage 1 area. Nine plots out of 121 were located on the road or in permanently wet areas and were ignored. An additional five plots were removed because their site synecological coordinates were outliers. A summary of site coordinates for the remaining 107 plots shows similar means for M and N of just under 2.5 with values ranging roughly from 2.0 to 3.0. Nutrients had slightly greater range and standard deviation than moisture.

Isotropic variograms were calculated, using a minimum distance of 45 m and a maximum of 500 m (75% of the largest distance.) Models were fit to the experimental variograms (figure 3). Model and parameters were estimated visually and provided excellent fits with  $R^2$  of .943 for M and .985 for N. The linear model was fit to M, the spherical model to N. The experimental variogram

values for stage 2 N data progressively increased with distance and then showed a systematic decrease, known as a hole effect, beyond the range of spatial dependence. Since variograms were plotted beyond half the maximum separation distance, the apparent hole effect is likely an edge effect. The models are fitted for values up to the range. In addition, anisotropy was examined. However, there were not enough data to accurately estimate apparently mild directional influences.

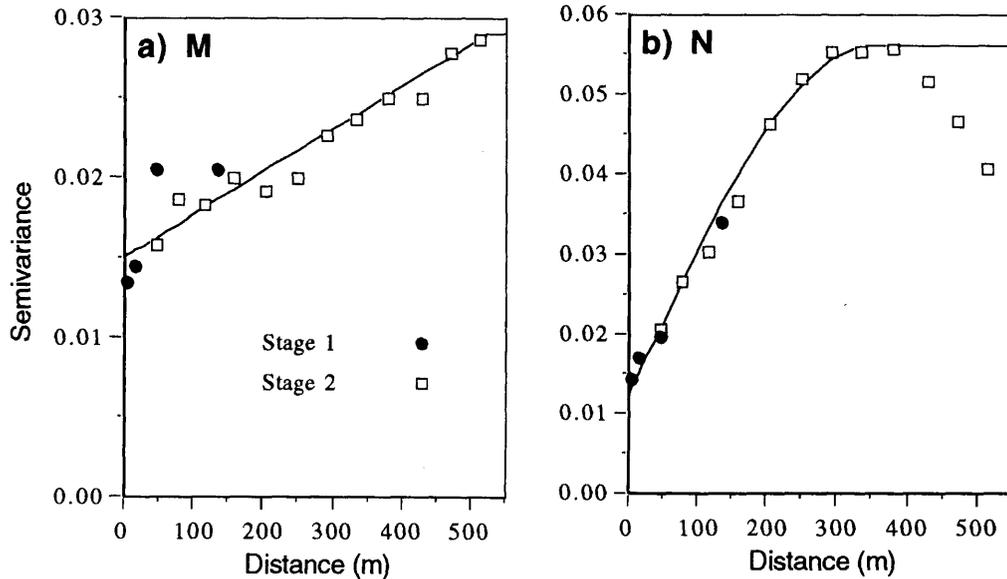


Figure 3.—Experimental and fitted variograms for site synecological coordinates: (a) Moisture and (b) Nutrients.

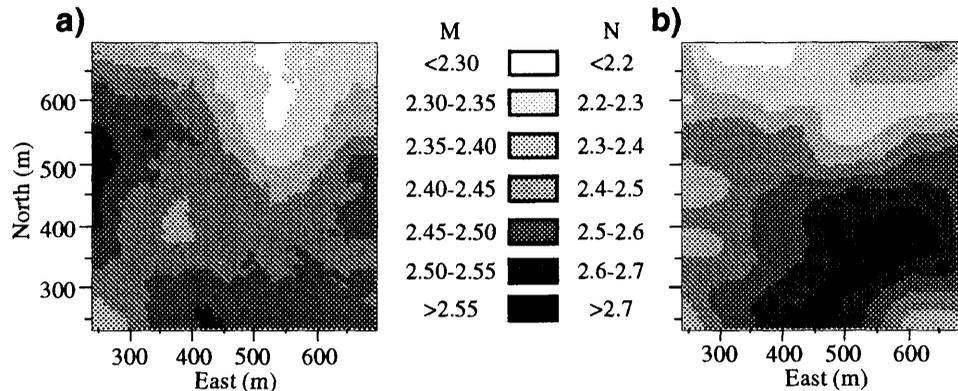
The supplemental information about the spatial variability at close distances confirms that there is at most a moderate nugget effect (i.e., random variation or spatial dependence at less than 5 m) in the experimental variograms, which ranges from 52% (for M) to 22% (for N) of the total sample variability. That signifies that 48 to 78% of the total variability is spatially dependent. If we assume small measurement errors, the large relative nugget for M appears to be the result of very closely spaced variation. N's range is 350 m and M's range is 525 m, indicating that our choice to resample on a 45-m grid was appropriate. Results of the cross validation are summarized (table 1). Ordinary kriging was performed on the grid points using the 16 nearest neighbors within a radius of 130 m. Cross validation statistics for median absolute error indicate that 50% of the observations will be within .113 units (for M) and .111 units (for N) of the true value.

Table 1.—Stage 2 cross validation statistics for Moisture (M) and Nutrients (N).

Factor	Mean Error	Median Abs Error	MSE
M	.001	.113	.018
N	.000	.111	.024

## SITE SYNECOLOGICAL COORDINATE MAPPING

We used the cross-validated variogram models to estimate values for M and N by applying ordinary block kriging. Using the previous search procedure, we produced contour maps (figure 4). Maps of the synecological coordinates were compared with topographic relief in the study area. No obvious relations with elevation were found. The nutrients contour map shows the presence of a large, homogeneous region to the southeast. Moisture seems more influenced by random or micro-relief variation, thus producing more irregular contours than nutrients.

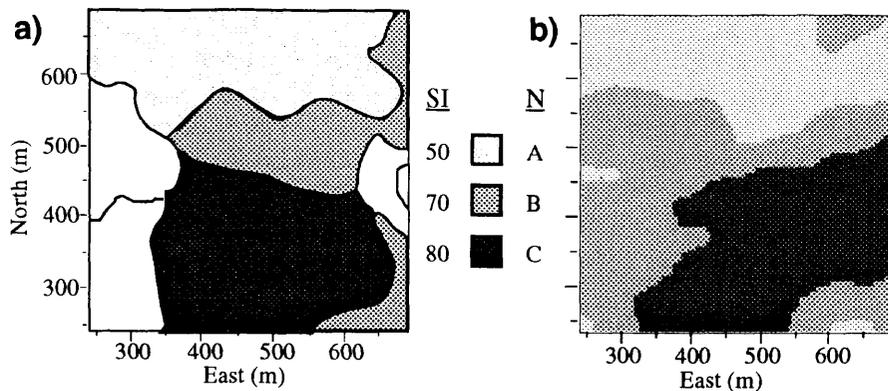


**Figure 4.—Contour maps of kriged estimates of site synecological coordinates for stage 2 data: (a) Moisture and (b) Nutrients.**

We looked for similarities between the spatial distribution of each coordinate and the inventoried SI. The spatial patterns of kriged N appear to yield the best relationship to the three SI classes (figure 5a). By plotting contours at 2.35 and 2.55 for N, we delineated three areas with  $A < 2.35 < B < 2.55 < C$  (figure 5b), which correspond roughly to areas with  $SI = 50, 70, \text{ and } 80 \text{ ft}$  (figure 5a). Estimated N coordinates in these upland aspen stands separated by .20 units may account for SI changes of 10 to 20 ft. Comparing figure 5a with 5b gives us reason to support the belief that vegetative-based nutrient estimates may be useful in depicting areas of differing site quality.

## DISCUSSION

Spatial dependence is clearly evident, which allows for mapping site synecological coordinates. The N coordinate showed greater spatial dependence than the M coordinate, which showed a relative nugget effect of roughly 50%. Viewed from another perspective, the variogram for M reveals two very different



**Figure 5—Comparison of site quality areas (a) as delineated by inventory site index classes and (b) as estimated by low (A), moderate (B), and high (C) Nutrient classes.**

and yet equally strong aspects to the variability: 50% is due to measurement error and very short-distance variability (less than 5 m), and 50% is due to long-range spatial dependence. For the N coordinate, the relative magnitude of the variability due to the long-range spatial dependence has increased and the magnitude of the micro-site variability has decreased.

Both coordinates showed ranges of at least 350 m, which is far greater than the usual limits of spatial dependence in soil studies. Spatial ranges for various soil properties range from 1 to 2 m (Lechowicz and Bell 1991, Jackson and Caldwell 1993) up to 15 to 40 m (Oliver and Webster 1991, Palmer 1990). The variation of soils is almost always more noisy or random than other environmental variables such as landform, climate, or geology, where long-range effects often dominate (Burrough 1983). The high proportion of short-range sources of variation in soil is caused by very localized biological activity, weathering, and small differences in relief due to glacial action, erosion, and deposition (Burrough 1983). Because of its large relative nugget, M appears more influenced by localized sources of variation than N. The large ranges and the large-scale patterns evident in figure 4 suggest that broader trends are operating for both factors. Synecological coordinates have reduced the complexity of the spatial variation so that micro-site variation in soil properties no longer dominates, but rather broader trends are revealed, such as the relationship of N coordinates to overstory productivity.

#### ACKNOWLEDGMENTS

Dr. Cindy Johnson-Groh, Janna Goerdts, and Marshall Kaiser of the College of St. Scholastica collected the plant lists. The St. Louis County Land Department generously allowed access to its property, helped locate a suitable study site (Tom Zeisler and Mark Weber), and provided GPS coordinates for the starting point (Glenn Berts).

## REFERENCES

- Bakuzis, E. V. 1959. Synecological coordinates in forest classification and in reproduction studies. Ph.D. Diss., Univ. of Minnesota, St. Paul. 244 p.
- Bakuzis, E. V., and Kurmis, V. 1978. Provisional list of synecological coordinates and selected ecographs of forest and other plant species in Minnesota. Univ. of Minnesota, Dept. of Forestry. Staff Pap. Ser. No. 5. 31 p.
- Brand, G. J. 1985. Environmental indices for common Michigan trees and shrubs. USDA For. Serv. Res. Pap. NC-261. 5 p.
- Burgess, T. M., and R. Webster. 1980. Optimal interpolation and isarithmic mapping of soil properties. II. Block kriging. *J. Soil Sci.* 31: 333-341.
- Burrough, P. A. 1983. Multiscale sources of spatial variation in soil. I. The application of fractal concepts to nested levels of soil variation. *J. Soil Sci.* 34: 577-597.
- Fortin, M., P. Drapeau, and P. Legendre. 1989. Spatial autocorrelation and sampling design in plant ecology. *Vegetatio* 83: 209-222.
- Gutiérrez-Espeleta, E. E. 1991. Tropical forest site quality assessment: an approximation in Costa Rica. Ph.D. Diss., Iowa State Univ., Ames. 138 p.
- Jackson, R. B., and M. M. Caldwell. 1993. The scale of nutrient heterogeneity around individual plants and its quantification with geostatistics. *Ecology* 74(2): 612-614.
- Lechowicz, M. J., and G. Bell. 1991. The ecology and genetics of fitness in forest plants. II. Microspatial heterogeneity of the edaphic environment. *J. Ecol.* 79: 687-696.
- Nimerfro, K., and Brand, G. 1993. The microcomputer scientific software series 7: data recorder program for storing plant lists and calculating synecological coordinates. USDA For. Serv. Gen Tech. Rep. NC-157. 20 p.
- Oliver, M. A., and R. Webster. 1986. Combining nested and linear sampling for determining the scale and form of spatial variation of regionalized variables. *Geo. Anal.* 18: 227-242.
- Oliver, M. A., and R. Webster. 1991. How geostatistics can help you. *Soil Use and Mgmt.* 7: 206-217.
- Palmer, M. W. 1990. Spatial scale and patterns of species-environment relationships in hardwood forest of the North Carolina piedmont. *Coenoses* 5: 79-87.
- Webster, R., and T. M. Burgess. 1983. Spatial variation in soil and the role of kriging. *Agr. Water Mgt.* 6: 111-122.

## BIOGRAPHICAL SKETCH

Margaret R. Holdaway is a Mathematical Statistician, North Central Forest Experiment Station, St. Paul, MN. She has been a member of the Station's Forest Inventory and Analysis Unit since 1995 (previously with the Modeling Unit) and holds an M.S. in Biometry from the University of Minnesota. Marge's areas of interest are spatial statistics, forest growth modeling, and climatic applications.

Gary J. Brand is Research Forester, North Central Forest Experiment Station, St. Paul, MN. He has been a member of the Station's Forest Inventory and Analysis Unit since 1995 (previously with the Modeling Unit) and holds an M.F. in Forestry from the University of Minnesota. Gary is interested in the relations between synecological coordinates, ecological classifications, and forest stand dynamics.

# On the Robustness of Data Assimilation Methods in Air Pollution Modeling

X.F. Zhang<sup>1</sup>

A.W. Heemink<sup>1</sup>

---

## Abstract

---

The performance robustness properties of two data assimilation methods, Kriging and Kalman filtering presented for the estimation of air pollution are investigated under perturbations of noise uncertainty and model parameter uncertainty. The simulation results are used to illustrate the advantages of Kalman filtering approach under various circumstances.

---

## INTRODUCTION

In Zhang *et. al* (1995) two data assimilation methods, a Kriging approach and Kalman filtering are developed for the estimation of two dimensional air pollution problem. The deterministic system model and the statistics of system noise and measurement noise are assumed to be known exactly in Zhang *et. al* (1995). However, such an assumption is often unrealistic in applications, and noticeable deterioration of system performance may be caused by model errors and noise uncertainty.

In this paper we investigate the performance robustness properties of the two data assimilation methods proposed in Zhang *et. al* (1995). The effects of both noise uncertainty and parameter uncertainty are examined. In Section 2, the stochastic system model is described, and in Section 3 the two data assimilation methods are briefly reviewed. In Section 4, the influence of noise uncertainty on the Kalman filtering approach is analysed, and a strategy to minimize the worst case performance is proposed. Then, in Section 5, simulation results are presented, showing the advantages of the proposed robust Kalman filter design method compared with the conventional Kalman filter design and the Kriging approach at present of noise uncertainty. Furthermore, the performance robustness is checked with respect to uncertain transport speed of air pollution. Finally, Conclusion is given in Section 6.

---

<sup>1</sup>Department of Technical Mathematics and Informatics, Delft University of Technology, Mekelweg 4, P.O. Box 5031, 2600 GA Delft, The Netherlands

## THE STOCHASTIC DYNAMIC AIR POLLUTION MODEL

The transport phenomena of air pollutant in the environment are usually described by an advection-diffusion equation

$$\frac{\partial c}{\partial t} = -\nabla \cdot [\mathbf{u}c - \nabla \cdot Dc] \quad (1)$$

where  $c$  is the concentration of the pollutant,  $t$  is time,  $\mathbf{u}$  is the velocity field,  $D$  is the dispersion tensor including molecular diffusion and dispersion. By introducing a numerical scheme (Heemink 1990, Van Eijkeren 1993) Eq. (1) can be represented by a discrete time state space model.

$$\mathbf{x}(k) = \Phi \mathbf{x}(k-1) + \Gamma \mathbf{w}(k) \quad k = 1, 2, \dots \quad (2)$$

$$\mathbf{z}(k) = H \mathbf{x}(k) + \mathbf{v}(k) \quad (3)$$

where  $\mathbf{x}(k)$  is an  $n$ -vector state process taking corresponding values at model grid point at time instant  $k$ ,  $n$  is the total number of model grid points at the pollutant concentration field, and  $\Phi$  is an  $n \times n$  dimension transient matrix obtained by the numerical scheme discretizing the advection-diffusion Eq. (1),  $\Gamma$  is an  $n \times q$  noise input matrix that interpolates  $q$  dimensional system noise  $\mathbf{w}(k)$  at  $n$  model grid points,  $\mathbf{z}(k)$  is an  $m$ -vector measurement process, with measurement noise sequence  $\mathbf{v}(k)$ .  $\Phi$ ,  $\Gamma$  and  $H$  are constant matrices,  $\mathbf{w}(k)$  and  $\mathbf{v}(k)$  and the initial state  $\mathbf{x}(0)$  are assumed to be multiply independent and satisfied the following conditions:

$$E\{\mathbf{w}(k)\mathbf{w}'(k)\} = Q\delta_{jk}, E\{\mathbf{v}(k)\mathbf{v}'(k)\} = R\delta_{jk} \quad (4)$$

$$E\{\mathbf{x}(0)\} = \bar{\mathbf{x}}(0), E\{[\mathbf{x}(0) - \bar{\mathbf{x}}(0)][\mathbf{x}(0) - \bar{\mathbf{x}}(0)]'\} = P_0 \quad (5)$$

where  $Q \geq 0$  is a positive semidefinite constant matrix,  $R > 0$  is a positive definite constant matrix,  $\delta_{jk} = 1$  if  $j = k$  and  $\delta_{jk} = 0$  if  $j \neq k$ .

## DATA ASSIMILATION METHODS

### The Kriging Approach

Kriging is a method for the optimal estimation of a quantity  $y$  at an arbitrary location that has not been measured by using measurements taken at surrounding points. By changing the position of the point of estimation, it is possible to estimate the whole field of spatial variables. the estimation of the pollutants concentration value  $x$  at  $i$ th grid point of  $G_m$  and time instant  $k$  is given by:

$$\begin{aligned} \hat{\mathbf{x}}_i(k|k) &= \Phi(i, :) \hat{\mathbf{x}}(k-1|k-1) \\ &+ \sum_{j=1}^s K(i, j) [\mathbf{z}_j(k) - H(j, :) \Phi \hat{\mathbf{x}}(k-1|k-1)] \\ & \quad i = 1, 2, \dots, n \end{aligned} \quad (6)$$

where  $s$  is the total number of measurement stations. To determine the optimal choices of  $K(i, 1), K(i, 2), \dots, K(i, s)$  of minimizing the estimation variance for an arbitrary  $i$ th grid point of  $G_m$ . We get (Zhang et. al 1995)

$$\sum_{l=1}^s (\gamma(h_{jl}) - R_{jl})K(i, l) + \mu = \gamma(h_{ji}) \quad j = 1, \dots, s$$

$$\sum_{j=1}^s K(i, j) = 1 \quad (7)$$

where  $\gamma_{j,l}$  denotes the value of the semi-variogram  $\gamma(h_{j,l})$  of the deviation field.  $R_{j,l}$  denotes the value of the known covariance of measurements noise. The spatial correlation in the form of a semi-variogram is automatically estimated based on the available measurement data at each time instant. The set of  $K(i, j), j = 1, \dots, s$  is obtained by Eq. (7) and then applied to Eq. (6). This produces an unbiased estimation with minimum variance of estimation error. The updated picture of the whole concentration field is obtained by adjusting  $\mathbf{x}_i$  throughout every grid point of the concentration field using Eq. (6).

### The Kalman Filter Approach

Since the linear discrete-time system Eq. (2) and Eq. (3) is time-invariant, the Chandrasekhar-type algorithm (Morf *et al* 1974) can be employed to get a steady state Kalman filter as follows:

$$\hat{\mathbf{x}}(k|k) = \Phi \hat{\mathbf{x}}(k|k-1) + K[z(k) - H\hat{\mathbf{x}}(k|k-1)] \quad (8)$$

$$\hat{\mathbf{x}}(k|k-1) = \Phi \hat{\mathbf{x}}(k-1|k-1) \quad (9)$$

where the gain matrix can be obtained by (Zhang *et. al* 1995):

$$G(k+1) = G(k) + Y(k)L(k)Y(k)'H' \quad (10)$$

$$R^e(k+1) = R^e(k) + HY(k)L(k)Y(k)'H' \quad (11)$$

$$K(k+1) = G(k+1)R^e(k+1)^{-1} \quad (12)$$

$$S(k+1) = Y(k) - K(k+1)HY(k) \quad (13)$$

$$L(k+1) = L(k) + L(k)Y(k)'H'R^e(k)^{-1}HY(k)L(k) \quad (14)$$

$$Y(k+1) = \Phi S(k+1) \quad (15)$$

with initial condition:

$$Y(1) = \Gamma, G(0) = 0, R^e(0) = R, L(0) = Q \quad (16)$$

The equations from Eq. (10) to Eq. (15) are iterated until:

$$\|K(k+1) - K(k)\| < \theta \|K(k)\| \quad (17)$$

where  $\theta$  is determined by prespecified accuracy. Under the assumption that the dimension of system noise process  $q$  is much less than the dimension of system states  $n$ , i.e,  $q \ll n$ , the Chandrasekhar-type algorithm provides a significant computation reduction (Heemink 1988,1990).

## A ROBUST FILTER DESIGN METHOD UNDER PERTURBATIONS OF NOISE UNCERTAINTY

In general, it is a hard task to get suitable  $Q$  and  $R$  matrices which fit best the stochastic system model Eq. (2) with various observations. We suppose that the real steady state covariances of system noise and measurement noise are uncertain, and they can be described by:  $Q = Q_0 + \Delta Q$ ,  $R = R_0 + \Delta R$ . Further, we suppose that all information on  $\Delta Q$  and  $\Delta R$  is that the norm of  $\Delta Q$  and  $\Delta R$  are bounded by:  $\|\Delta Q\| \leq \alpha$ ,  $\|\Delta R\| \leq \beta$ . As the optimal gain matrix  $K$  of the Kalman filter is a function of  $Q$ , and  $R$ , the  $K$  obtained by using  $Q_0$  and  $R_0$  is in general no longer optimal due to the existence of uncertainties in  $Q$  and  $R$  matrices. First, we consider the prediction-type Kalman filter with gain matrix  $K_p$ . Let  $e(k+1) = x(k) - \hat{x}(k|k-1)$  the actual estimation covariance denoted by  $N^0(k+1|k)$  is determined by

$$\begin{aligned} N^0(k+1|k) &= E\{e(k+1|k)e'(k+1|k)\} \\ &= (\Phi - K_p H)N^0(k|k-1)(\Phi - K_p H)' \\ &\quad + \Gamma(Q_0 + \Delta Q)\Gamma' + K_p(R_0 + \Delta R)K_p' \end{aligned} \quad (18)$$

Sangsuk-Iam (1990) prove that as  $k \rightarrow \infty$  Eq. (18) always converges to some constant  $N^0$  which satisfies the following algebraic Riccati equation

$$N^0 = (\Phi - K_p H)N^0(\Phi - K_p H)' + \Gamma(Q_0 + \Delta Q)\Gamma' + K_p(R_0 + \Delta R)K_p' \quad (19)$$

From Eq. (19) we notice that the actual estimation covariance  $N^0$  is a monotonic increasing function of the actual noise covariances  $Q_0 + \Delta Q$  and  $R_0 + \Delta R$ . With  $\bar{Q} := \Delta Q_0 + \alpha I$ ,  $\bar{R} := \Delta R_0 + \beta I$  where  $I$  is an identity matrix, we have  $\bar{Q} \geq Q_0 + \Delta Q$ ,  $\bar{R} \geq R_0 + \Delta R$ . It is a reasonable choice to take the worst case covariance  $\bar{Q}$  and  $\bar{R}$  to design the Kalman filter which minimizes the upper bound of Kalman filter performance, i.e

$$N(K_p^*, \bar{Q}, \bar{R}) \leq N(K_p, \bar{Q}, \bar{R}) \quad (20)$$

where  $K_p^*$  is the optimal gain matrix of the Kalman filter designed for the fixed pair of worst case noise covariances  $Q_0 + \alpha I$  and  $R_0 + \beta I$ , and  $K_p$  could be any other gain matrix. We have the following saddle point inequality:

$$N(K_p^*, Q_0 + \Delta Q, R_0 + \Delta R) \leq N(K_p^*, \bar{Q}, \bar{R}) \leq N(K_p, \bar{Q}, \bar{R}) \quad (21)$$

Now, we consider the current estimation-type Kalman filter with gain matrix  $K_e$ . From Eq. (2), Eq. (3), and Eq. (8) we have the estimation error covariance of the form:

$$\begin{aligned} P(k) &= (I - K_e H)\Phi P(k-1)\Phi'(I - K_e H)' \\ &\quad + (I - K_e H)\Gamma Q \Gamma'(I - K_e H)' + K_e R K_e' \end{aligned} \quad (22)$$

Obviously, for any determined  $K_e$ , the actual estimation covariance  $P(k)$  is also a monotonic increasing function of  $Q + \Delta Q$ , and  $R + \Delta R$ . We have following saddle point inequality

$$P(K_e^*, Q_0 + \Delta Q, R_0 + \Delta R) \leq P(K_e^*, \bar{Q}, \bar{R}) \leq N(K_e, \bar{Q}, \bar{R}) \quad (23)$$

where  $K_e^*$  is the optimal gain matrix of the current estimation-type Kalman filter designed using the fixed pair of worst case noise covariances  $Q_0 + \alpha I$  and  $R_0 + \beta I$ , and  $K_e$  could be any other gain matrix. From above discussion we can see that the selection of worst case noise covariances is a strategy to design a robust Kalman filter (either prediction-type or current estimation-type) which minimizes the upper bound of the Kalman filter performance.

### SIMULATION STUDY

The simulation studies are aimed to examine the performance robustness of the two data assimilation methods with respect to noise uncertainty and model parameter uncertainty. The experimental region is a cycle area bounded in  $\Omega$  (see figure 1).

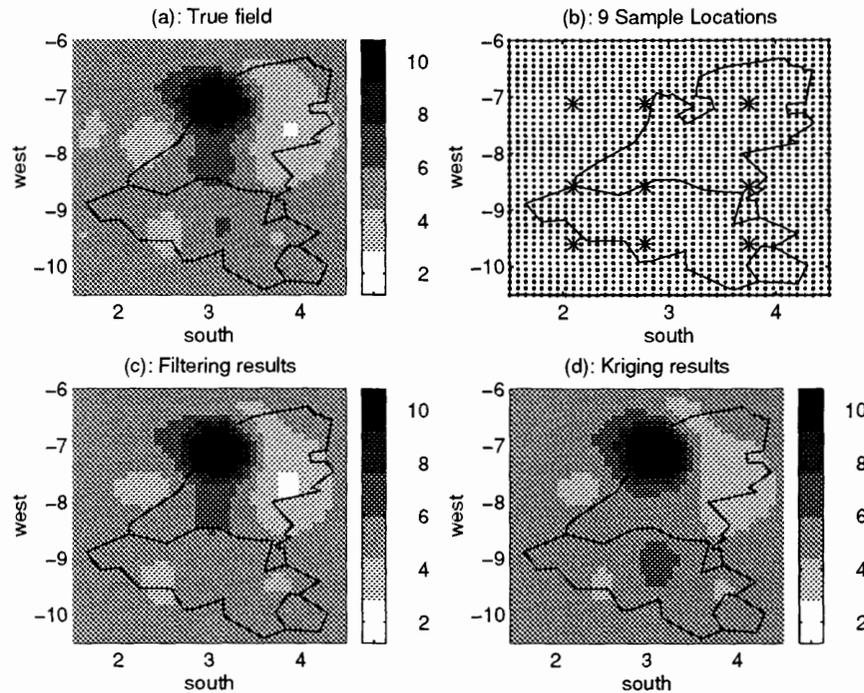


Figure 1: (a): The concentration field at  $t=64$ ; (b): “\*” Observations, “.” Model grid  $G_m$ ; (c), (d): The corresponding results of the Filtering and the Kriging approaches without model parameter and noise uncertainty.

The concentration of the pollutant is denoted by  $\mathbf{x}(\boldsymbol{\xi}, t)$  where  $\boldsymbol{\xi}$  is the coordinate  $\boldsymbol{\xi} = (\underline{x}, \underline{y})' \in \Omega$ . The true field is a stochastic process described by Eq. (2). The deterministic part of this stochastic process is driven by the advection-diffusion equation Eq (1) with a circular velocity  $\mathbf{u} = u((\underline{x}, \underline{y})') = 2\pi(-\underline{y}, \underline{x})'$  by using the numerical approximation described in Heemink (1990) and Van Eijkeren (1993). The initial condition of Eq. (1) consists of a constant background field and a small cosine-squared shape cone interpreted as a high density concentration area of local pollution. In our experiments, the simulated true field at  $t = 64$  is shown in figure 1, (a). The regular model grid  $G_m$  is shown in figure 1, (b). The system noise  $\mathbf{w}$  is generated by  $\mathbf{w} = (w_1, \dots, w_q)'$  with zero mean and constant covariance  $Q_0(i, j) = 2 \times 10^{-2} e^{-\frac{1}{0.06}[(\underline{x}_i - \underline{x}_j)^2 + (\underline{y}_i - \underline{y}_j)^2]}$  and the measurement noise  $\mathbf{v}$  is generated by  $R_0(i, j) = 1 \times 10^{-2} e^{-\frac{1}{0.01}[(\underline{x}_i - \underline{x}_j)^2 + (\underline{y}_i - \underline{y}_j)^2]}$ . The location of 9 observation stations is also plotted in figure 1 (b). When the deterministic model, the statistics of system noise and measurement noise are exactly known the reasonable good results of both data assimilation approaches are plotted in figure 1 (c), (d).

### Experiment 1

We assume that the second moments of statistics of the system noise and measurement noise are uncertain. Let nominal filter use the optimal gain matrix corresponding to the nominal system noise covariance  $Q_0$  and measurement noise covariance  $R_0$ , and robust filter use the robust gain matrix which minimize the worst case of  $\bar{Q} = Q_0 + \alpha I$  and  $\bar{R} = R_0 + \beta I$  with  $\alpha = 1 \times 10^{-4}$  and  $\beta = 5 \times 10^{-5}$ . *First*  $Q = Q_0$  and  $R = R_0$ : The Root Mean Square (RMS) of the nominal filter and robust filter are plotted in (a) of figure 2. It can be seen that the RMS corresponding to robust filter is only slight higher than that of nominal filter which is the optimal gain matrix in this case. *Secondly*  $Q = Q_0 + \alpha I$  and  $R = R_0 + \beta I$ : The RMS of the nominal filter and robust filter are plotted in (b) of figure 2. We can see that the RMS of robust filter is now much lower than that of nominal filter. It is obvious from above simulation results the robust filter which minimize the worst case covariance of the estimation error makes the performance of the Kalman filter much better from the point of view of robustness. To compare the two data assimilation methods the corresponding RMS of the Kriging approach are also plotted in figure 2 (a), (b). It can be seen that Kalman filter approach can provide more accurate result than The Kriging approach.

### Experiment 2

We suppose that the Kalman filter is perturbed by uncertain deterministic system model. More exactly, it is assumed that the transport speed  $\mathbf{u}$  which is a key model parameter is uncertain. Let the Kalman filter be designed by  $\mathbf{u}_0 = (u_x, u_y) = 2\pi(-y, x)$ , and let the true concentration field is run with

$\mathbf{u} = \mathbf{u}_0 + \Delta\mathbf{u}$ .  $\Delta\mathbf{u} = -0.1\mathbf{u}_0$  i.e. 10% deviation. The true concentration field at  $t = 64$  is plotted in (c) of figure 2. The corresponding estimated field of the Kalman filter approach and the Kriging approach are plotted in (d) and (e) of figure 2, respectively. The RMS of the data assimilation procedures are plotted in (f) of figure 2. It is clear in this case that the performance robustness of the Kalman filter approach to tolerate model parameter uncertainty is much better than that of the Kriging approach.

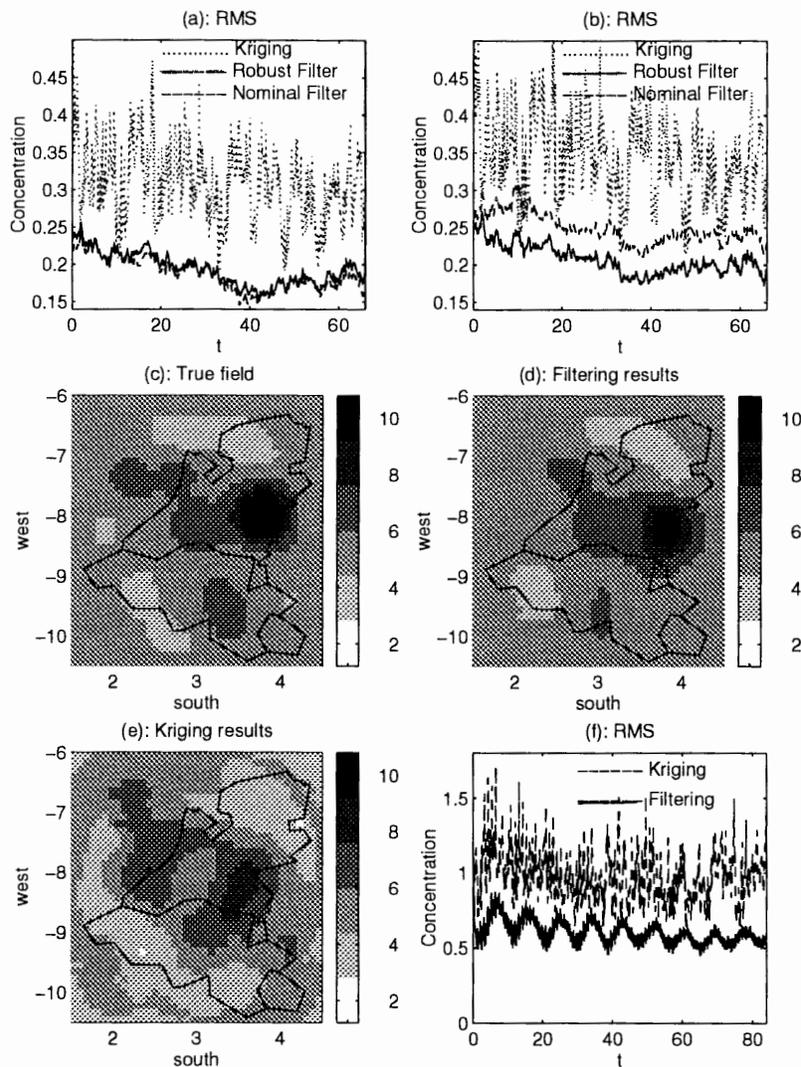


Figure 2: (a) and (b): The RMS of three methods with noise uncertainty. (c): Truth field at  $t=64$ ; (d) and (e): The corresponding results of Filtering and Kriging approaches with 10% model parameter uncertainty; (f): RMS of both two approaches.

## CONCLUSION

In this paper, the performance robustness of two data assimilation methods under the perturbations of noise uncertainty and parameter uncertainty is investigated in two dimensional air pollution problem. The experimental results show the advantage of a robust Kalman filter design method with respect to noise uncertainty. Moreover, it is demonstrated in the simulation studies that the Kalman filter approach has much stronger performance robustness to tolerate model parameter uncertainty than the Kriging approach. Our current research focuses on applying Kalman filtering to “ the evaluation of CH<sub>4</sub> budget in Europe ” in cooperation with RIVM air quality laboratory.

## ACKNOWLEDGMENTS

This work has been carried out in cooperation with and with financial support from the RIVM.

## REFERENCES

- Heemink, A.W., 1988. Two-Dimensional Shallow Water Flow Identification. *Applied Mat. Mod.*, 109-118.
- Heemink, A.W., 1990. Identification of Wind Stress on Shallow Water Flow Surfaces by Optimal Smoothing. *Stochastic Hydrol. Hydraul.* Vol. 4, 105-119.
- Morf, M., Sidhu, S.S., Kailath, T., 1974. Some New Algorithms for Recursive Estimation in Constant, Linear, Discrete-Time system. *Trans. Autom. Control*, Vol. AC-19, NO. 4, 315-323.
- Sangsuk-Iam, S., 1990. Analysis of discrete-time Kalman filtering under incorrect noise covariance, *IEEE Transactions on Automatic Control*, Vol. 35, 1304-1309.
- Van Eijkeren, J.C.H., 1993. Backward Semi-Lagrangian Methods:an Adjoint Equation Method. In: *Numerical Methods for Advection - Diffusion Problems*, Vreugde hil C.B and Kora B Eds. Vieweg, Vol, 45, 215-241.
- Zhang, X.F., Van Eijkeren, J.C.H. and Heemink, A.W. 1995 Data Assimilation in Dynamic Environmental Pollution Modeling. RIVM report, No. 421503 006, Bilthoven, The Netherlands.

## BIOGRAPHICAL SKETCH

Zhang X.F. is doing a Ph.D. project on the development of data assimilation methods for the estimation of air pollution.

Heemink A.W. is a professor of faculty of technical mathematics and informatics, TU Delft.

# The Development of a GIS Information Quality Module

Jane Drummond <sup>1</sup>, Allan Brown <sup>2</sup>, Du Daosheng <sup>3</sup>, Corné van Elzakker <sup>4</sup>

---

**Abstract.**- In the early 1980s the authors first considered 1) the transfer of traditional mapmakers' expertise regarding spatial information quality into GIS and 2) ways of determining GIS generated information's quality. Taking into account the now accepted principles that attribute, position, age, and topology quality and lineage all contribute to GIS generated information, the framework for our work was that as a GIS generated new information an associated uncertainty subsystem should in parallel generate information on the quality of that information and that the two sets of information should be displayed simultaneously to better the GIS user's decision making. Now an extant (though incomplete) entity the subsystem consists of appropriate means to determine and store the quality of spatial data and the processing model used to generate information, means for determining the generated information quality (variance propagation, fuzzy sub-set theory, etc.) and means for visualizing that quality. These are integrated in an Uncertainty Subsystem called PUIUS. The subsystem has been implemented in ILWIS. This paper will review the subsystem (and another), summarize some applications and point the way to further work.

---

## INTRODUCTION

Maintaining high standards has long concerned land surveyors, photogrammetrists and cartographers, but the Geographic Information System (GIS) community seemed only alerted to the problem of information quality in the early 1980's. By the end of the 1980's those in GIS had accepted the principles that: quality of attribute and position; information age; topological consistency; and lineage all contributed to the quality of any GIS generated information. We first started, formally, to consider 1) the transfer of traditional mapmakers' expertise regarding spatial information quality in the

---

<sup>1</sup> Jane Drummond, Topographic Science Section, Glasgow University, UK

<sup>2</sup> Allan Brown, Department of Geoinformatics, ITC, Enschede, The Netherlands

<sup>3</sup> Du Daosheng, Professor -Wuhan Technical University of Surveying and Mapping, Wuhan, PR China

<sup>4</sup> Corné van Elzakker, Department of Geoinformatics, ITC, Enschede, The Netherlands

GIS environment and 2) ways to determine the quality of GIS generated information, in the mid-eighties.

Taking into account the principles presented in the first paragraph, our study's framework was that while a GIS generated new information an associated uncertainty subsystem should in parallel generate information on the quality of that new information and that the two sets of information should (if possible) be displayed simultaneously to better the GIS user's decision making. The subsystem itself would consist of appropriate means to determine (knowledge base, user queries, etc.) and store the quality of spatial data, to store the quality of the processing model used to generate information, the means for determining the generated information quality (variance propagation, fuzzy sub-set theory, etc.) and the means (2D, 3D) for visualizing that quality. These were integrated in an Uncertainty Subsystem called PUIUS (Prototype User Interface for an Uncertainty Subsystem). The subsystem has been implemented in ILWIS, a PC based GIS developed and marketed by the International Institute for Aerospace Survey and Earth Sciences (ITC) in The Netherlands. This paper will review this subsystem.

The review will take the form of a summary of projects carried out by the group which lead to the development of PUIUS, and an investigation of a recent version of another off-the-shelf GIS package which is now marketed with some information quality processing tools. Thus the five remaining sections of this paper will deal with i) early ideas; ii) the XGIS project; iii) the PUIUS project; iv) the Idrisi's information quality processing tools; v) and conclusions.

## **EARLY IDEAS**

In 1983 two of this paper's authors, working at the UK Experimental Cartography Unit, prepared a very straightforward presentation on digital spatial data quality for a Visiting Committee. The presentation showed boundary widths for polygons of various natural features as a function of the standard deviation (SD) of these boundaries' coordinates, and then the resulting overlaid polygons. Some visitors expressed surprise at the demonstrated imprecise nature of the digital spatial data we so expensively acquired. Although official British science seemed unready pay for recognisably imperfect digital spatial data (the two authors concerned relocated to the Netherlands and China), at about the same time Blakemore (1983) and Chrisman (1983) were much more successfully convincing European and North American academics that the imperfections of digital spatial data represented a respectable research area. The presentation's polygons represented areas of (assumed) uniform slope, soil class, rainfall and temperature and were overlaid and used for crop suitability mapping

(ODA,1983). At the time, and working in the vector environment, to determine the precision of such contributing polygons' boundaries and then produce overlaid polygons with boundary line widths proportional to their (e.g.) precision was cumbersome. The same problem has now been much more elegantly addressed in the raster environment of PUIUS (see below).

The user of a GIS is not likely to be interested in the quality of the data *per se*, but in the quality of the generated information. In the ODA suitability mapping project, the input sets of polygons each had a centroid - as had each polygon resulting from overlay. Standard databasing facilities permit such centroids to have attributes referring to data quality and can be easily processed. Such was part of the thinking behind much subsequent work in the group (Drummond,1987). The group understood a GIS to be both:

- (i) a system which used processing models to generate new information from existing data; and,
- (ii) a spatial data archive which allowed one to retrieve and display stored data.

In the first case the **new information** will be visualized, but simultaneously (or by toggling) its **quality** should be visualised. Error propagation models are directly related to information generation models and information generation models can be mathematical (e.g. computation of aspect; USLE) or logical (e.g. crop suitability; Community Tax determination). Thus if a mathematical information processing model gives us  $X = f(a,b,c)$  then  $\sigma X = f(\sigma a, \sigma b, \sigma c)$ , where  $\sigma$  represents Standard Deviation. Likewise if a logical processing model gives us  $Y = f(l,m,n)$  then  $pY = f(pl, pm, pn)$  where  $p$  represents a certainty statistic such as Probability or Certainty Factor. However the processing model itself may be in error, thus however good the data are, the generated information may be poor. In the case of a mathematical model its error can be handled as noise and subsequently processed by  $\sigma X = f(\sigma a, \sigma b, \sigma c, \sigma n)$  where  $n$  represents noise. In the case of a logical model, the model itself may have a certainty statistic associated with it which can be processed by  $pY = f(pl, pm, pn, po)$  with  $po$  representing probability or CF of the model holding.

In the second case above a GIS might, e.g., be queried to produce a map to show the position of a utility line through a neighbourhood. The users of the map might be concerned with the accuracy of depiction of the utility's location. If the precision of the procedure used to survey in the utility was recorded in the attribute tables of the GIS, then this information could be extracted and used to control the width (e.g.  $3\sigma$  allowing >99.9% probability of finding the utility) of the utility's mapping. In this case no processing model is used to generate new information.

Following our understanding of GIS, a simple GIS using the merged capabilities of AutoCAD and dBASE (Drummond, 1990) was built. As well as more usual facilities, this GIS had an interface permitting interrogation of the

user to acquire quality information of each data set. Although not a very user-friendly interface (when compared to PUIUS described in a later section) it permitted the user to provide the quality information of each data item in  $\sigma$ , probability or natural language terms convertible to CF, as appropriate. Conversion (for example converting  $\sigma$  to probability using central limit theory when a continuous variable, such as rainfall, was reclassified into a discontinuous variable such as the rainfall class 300-400 mm a year) was carried out within the GIS. The acquired information was stored in the relevant attribute database table and processed using fuzzy sub-set reasoning to provide CF values for each crop suitability polygon (Drummond,1987). As the data concerned (i.e. for the crop suitability project) were all of a discontinuous type error propagation using fuzzy logic rather than variance propagation was appropriate - but variance propagation was supported. Simple maps showing the quality of generated suitability information, by polygon and with quality a function of hatching density, were produced.

FIGURE 1  
UNCERTAINTY SUBSYSTEM.  
The overall concept.

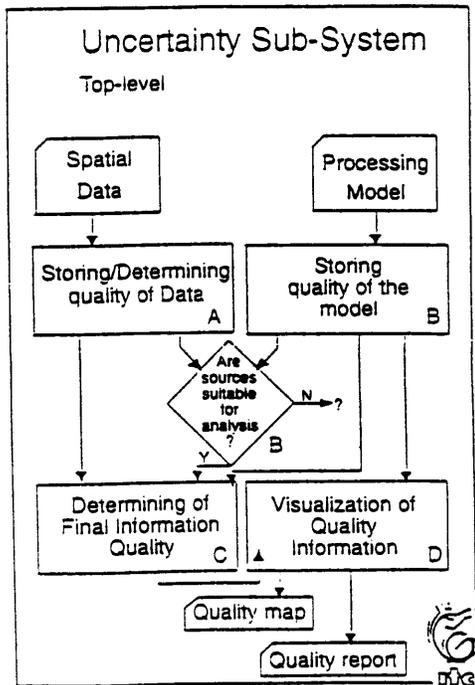
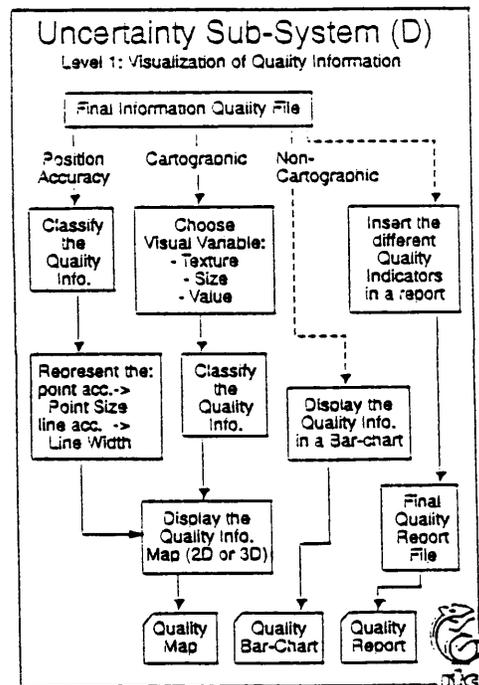


FIGURE 2  
UNCERTAINTY SUBSYSTEM  
Visualization of Information  
Quality



## THE XGIS PROJECT

The ITC was willing to partially support three MSc students (Ramlal,1991), (Fan,1992), (Oliveira,1992) to research data quality matters in their 'dissertation year' and fund a research associate (Barsoum,1995). This was referred to as the XGIS project and represented the development of an Uncertainty Subsystem - a subsystem of ILWIS (Ramlal et al., 1991). The overall concept of this subsystem is provided in FIGURE 1.

From FIGURE 1 it can be seen that the Uncertainty Subsystem was designed with four components. These dealt with the uncertainty of data (Storing/Determining quality of data), the uncertainty of processing models (Storing quality of the model), error propagation (Determining of Final Information Quality) and the visualisation of uncertainty (Visualisation of Information Quality). The visualisation was tackled first (Ramlal,1991), (Ramlal et al.,1992), (Van Elzakker et al. 1992). FIGURE 2 summarises this.

From FIGURE 2 it can be seen that the visual variables **texture**, **size** and **value** were considered appropriate for depicting data quality. The above mentioned work by Ramlal, Van Elzakker and Drummond resulted in maps of a type in which suitability was represented by the visual variable **colour** (e.g. green = most suitable to red = least suitable) and data quality by **value** (e.g. dark most certain to light least certain). Later Brown and Van Elzakker (Brown et al.,1993) experimented with the use of colour in the representation of categoric area information quality. They concluded that the simultaneous representation of attribute and quality is possible (by means of the **colour hue** and **colour saturation** respectively), for both screen display and printer output, for a limited number of categories and quality levels.

In tackling model quality (Fan,1992), Fan took on the task of developing Part B (FIGURE 1) of the Uncertainty Sub-System addressing (regression) model building within the GIS environment. By using Linear Least Squares Regression in model building, information became available on the performance (quality) of the model. This information was graphically displayed allowing the model developer to improve the model. Actual model development related to toxic cloud release was carried out. Fan established the concept of a model library in ILWIS where developed and pre-existing models could be stored along with relevant quality and other information.

De Oliveira addressed data quality (Oliveira,1992) (Storing/Determining Data Quality in FIGURE 1) and produced an appropriate module for the subsystem. The module is capable of automatically generating data quality parameters from related 'Ground Truth' data and user-supplied information. The algorithms, in the main, concerned positional and attribute accuracy assessment, working at two levels. The first level referred to overall accuracy parameters at the data set level (e.g theme) and the second represented accuracy parameters associated with individual database objects. It was a

requirement of De Oliveira's future employer (University of Aveiro) that concepts developed and implemented for ILWIS could also be implemented in Arc/Info to support an Emergency Response GIS. Error propagation for the particular information generation (or processing) models used (relating to chemical accidents in the form of toxic cloud releases) was developed.

The challenge of developing a complete error propagation module (i.e. Part C of the Uncertainty Sub-System: Determination of Final Information Quality - see FIGURE 1) has remained unaddressed in the XGIS project. Instead the remaining funds were used to support development of the user interface (within the PUIUS Project). User interface is an important issue if GIS users are to give serious consideration to matters of data and information quality.

### **THE PUIUS PROJECT**

In its original form ILWIS operated in the MS-DOS environment. With the decision to market ILWIS for WINDOWS from 1995 it was necessary to transfer XGIS into WINDOWS. In its prototype form and in the WINDOWS environment the ILWIS Uncertainty Subsystem has the name of PUIUS (Prototype Uncertainty Subsystem). It was hoped that the PUIUS project would complete the XGIS development, at least in prototype form. Referring to FIGURE 1 the transfer of parts B and D was completed, and of part C for discontinuous variables, but not continuous variables (Barsoum, 1995).

The prototype was developed using Visual Basic. The prototype deals with the assessment and representation of the quality of discontinuous (categorical) attribute data and resulting information referring to polygons. Three types of quality assessment are involved, namely fuzzy set theory using Certainty Factors (CFs), Bayesian probability theory and an integrated approach involving also the positional accuracy of polygon boundaries. In essence, the prototype is intended to provide a user-friendly graphical interface to assist ILWIS users to assess the quality of their data and information, and to visualize the results of the assessment using a wide variety of techniques.

Standard ILWIS functionality provides the established tools for examining data and analysing quality, e.g. confusion matrices, frequency graphs. PUIUS integrates these. The graphical user interface has a main menu to allow the user to opt to investigate data quality or model quality, and also to select an appropriate visualization technique. As already indicated, the data quality module has been partially implemented in the prototype, the model quality module not. The visualization module allows a wide choice of techniques, including tables, graphs, maps and oblique views: for example to perform a spatial analysis and simultaneously display the result. The associated quality information can be visualized by one of the available methods, or it can be accessed in numerical form by clicking on a polygon or pixel.

In the case of Certainty Factors, visualization was applied to the quality of a crop suitability map. As the default representation, four suitability classes, very suitable to not suitable, are represented by four hues (green, yellow, orange, red). The lower the CF, the greyer (i.e. lower saturation) the hues become. The user has control over hue choice and the rate of decreasing saturation. If desired, the visual variable **value** (lightness) can be used to represent uncertainty instead of, or as well as, **saturation**.

In the case of quality assessment using probabilities, PUIUS allows the adjacent display of results with or without using prior probability. As for the CF case described above, the visual variables **saturation** and/or **value** can be used to represent uncertainty. Also, for example for a land use map resulting from the classification of satellite data, blinking can be used to highlight particular pixels. Several criteria can be selected, e.g. those with classification accuracy above or below a specified probability threshold, those for which the classification changes if prior probability is applied.

### **IDRISI'S INFORMATION QUALITY PROCESSING TOOLS**

Apart from the above-reported work carried out in the ILWIS environment the only other off-the-shelf GIS which (as the authors are aware) appears to give consideration to matters of information quality is Idrisi (Idrisi,1995). Members of the group were able to examine this in 1995. This is a developing system so findings relate to a situation which may now have been superseded.

The investigation consisted of the execution of two different spatial information generation tasks using Idrisi and analysis of any uncertainty processing carried out during their execution. In the case of one test Idrisi handled the uncertainty in a near complete manner (the location of a new municipal dump - essentially a buffering operation) and in the case of the other (identification of agroclimatic zones based on computed environmental characteristics) some software development was needed (Du,1995).

Idrisi allows the insertion of either  $\sigma$  or probability values for spatial data. If these values have been inserted for some processing modules, error will be propagated. The user has to know what values to insert. We chose, as far as possible, to use data quality values which are standard for medium scale topographic data as supplied in PR China. There follow brief descriptions of Idrisi modules which address uncertainty. (These are discussed further in (Idrisi,1995) under the title "Uncertainty and Risk" and "Error Propagation Formulas" where some, perhaps over-, simplified formulae for variance propagation are presented. These formulae do not seem to be implemented directly, but having been presented the users can implement them themselves.)

PCLASS is an uncertainty handling version of a density slicing module. With PCLASS for some newly generated classes (e.g. <2000m, <4000m,

<6000m etc.) a probability of each element in the DTM falling in that new class is calculated, and stored in a displayable probability image, from user supplied information on the quality of the original DTM given as a standard deviation value. The central limit theorem is applied.

BAYES allows the user to exploit Bayesian Probability theory. The user can exploit not only the quality of the datasets being used in the analysis (e.g. the standard deviation of river or road coordinates processed through PCLASS to give a probability image) but also the quality of the processing model (e.g. that the particular rule for identifying suitable new landfill dump sites only works 54% of the time) and other (*a priori*) information (e.g. an older study identifying potential landfill dump sites in the same area). The usefulness of this tool is enhanced because in the absence of *a priori* information one can still exploit the module (although the procedure would no longer be strictly Bayesian) and at least use the information available on data quality.

FUZZY allows an image of uncertainty values to be generated. These uncertainty values are generated by one of three different types of membership function (unlike PCLASS which only uses central limit theorem). Subsequent processing of the FUZZY images must be performed using Idrisi's OVERLAY module, using the appropriate FUZZY Logic rules (i.e. that the minimum available certainty factor is assigned to the output image in the case of intersection, that the maximum certainty factor is assigned in the case of union, etc.). The assignment of certainty factors (via the different membership functions) can be based on the experience or knowledge of the user - rather than more objective information, such as standard deviation.

ERRMAT allows the generation of an error matrix (sometimes called a confusion matrix in Remote Sensing literature) showing the agreement between the image (e.g. landuse derived from the classification of a satellite image) and the 'truth' (perhaps obtained by a set of ground observations and used as check points).

MCE (Multi-Criteria Evaluation) supports information generation using several sets of input data and user input processing models. If standard deviation values (called RMS values in Idrisi) are stored in the documentation file of each input data set, then Idrisi will carry out a simplified variance propagation to give an estimate of output error, storing it in the relevant document file. The simplification is such that the results imply no variation in error across the resulting image, which is neither a helpful nor a valid assumption when the main purpose of uncertainty analysis is to guide the decision maker to geographic areas of greater certainty. A more rigorous approach has been implemented by the authors.

Other Idrisi modules contribute to the handling or understanding of uncertainty and include: REGRESS - for linear regression analysis; and RANDOM - for simulating error.

## CONCLUSIONS

The academic community investigating quality aspects of GIS is large. This group of authors has regarded the processing of uncertainty as a GIS sub-system. Others may regard it as a GIS analytical tool; as with many GIS analytical tools it may never be as well developed in an off-the-shelf system as in a system developed by specialists particularly interested in error. However this group has been characterised by individuals who believe that the standard (i.e. off-the-shelf tools) should be as good as possible and this is reflected in the fact that all developments have taken place in well-known environments (AutoCAD+dBASE; ILWIS; Arc/Info; Idrisi). Obviously development of error analysis tools must continue to take place without the constraints of off-the-shelf-systems, but it should always be possible to implement the resulting tools in such systems - if they are to have any effect in the 'real world'.

We have demonstrated that most components of the Uncertainty Sub-System can be implemented in ILWIS and the sub-system could thus become a standard product and marketed. Although Idrisi's information quality generation procedures are (also) incomplete, they are marketed. One can commend Clark University. By making users at least sometimes think about information quality such a GIS becomes an enhanced decision support tool.

## ACKNOWLEDGEMENTS

Professor Du Daosheng's fellowship at the Topographic Science Section, Glasgow University, in 1995, was funded by the Royal Society, 6 Carlton House Terrace, London. The efforts of our supervisees and colleagues, without whom much of the work referred to in this review would have remained undone, namely, Beshem Ramlal, Jorge de Oliveira, Fan Shien-ta, Elia Barsoum and Stephen McGinley are gratefully acknowledged.

## REFERENCES

- Barsoum, E.M., 1995 "A User Interface for Error Visualisation in ILWIS". report to the Division of Cartography, ITC, 1995.
- Blakemore, M. 1983, "Generalization and Error in Spatial Data Bases", Proceedings of AutoCarto 6, 1983, pp.313-322
- Brown A. and Elzakker, C.P.J.M van, 1993. "The use of colour in the cartographic representation of information quality generated by a GIS." In: P. Mesenburg (ed.), Procs. Vol 2. 16th International Cartographic Conference, Koln, 3-9 May 1993
- Chrisman, N.R. 1983, "The role of quality information in the long-term functioning of a GIS", Proceedings of AutoCarto 6, 1983, pp.303-312.
- Drummond, J., 1987 "A Framework for handling error in GIS", ITC Jo.1987/1

- Drummond, J., 1990 "Models and Data Quality handled in a dBASE/AUTOCAD GIS", in ITC Journal 1989 3/4
- Du Daosheng, 1995. "Investigation of information quality processing in an off-the-shelf GIS". Report prepared for the Royal Society of London at Topographic Science Section, Glasgow University.
- Fan, Shien-ta 1992 "Uncertainty Subsystem: Assessment of Model Quality", MSc Thesis, ITC 1992
- Idrisi for WINDOWS, Manual, 1995. Clark University Grad. School of Geog.
- ODA, 1983. "Dominica Land Use Planning Project". LRDC, Tolworth Tower, Surbiton, Surrey, KT6 7DY, UK. Report #127.
- Oliviera, J.P. de 1992 "Generation of Data Quality Parameters in a GIS", MSc Thesis, ITC 1992
- Ramlal, B and Drummond, J.E., 1992 "A Prototype Uncertainty Subsystem implemented in ITC's ILWIS PC based GIS and tested on a Dutch Land Reallotment Project", EGIS Conference Proceedings, Munich 1992
- Ramlal, Bhesem 1991 "Communicating Information Quality in a GIS Environment", MSc Thesis, ITC 1991
- Van Elzakker, C., Ramlal, B and Drummond, J.E., 1992 "The Visualisation of GIS Generated Information Quality", August 1992, Archives ISPRS Congress XVI, Commission IV

### **BIOGRAPHICAL SKETCH**

Jane Drummond is a staffmember of Glasgow University's Topographic Science Section. Prior to that she worked in ITC's Department of Geoinformatics, the UK NERC's Experimental Cartography Unit, UNB's Department of Geodesy and Geomatics Engineering, and Hunting Surveys.

Allan Brown is a staffmember of ITC's Department of Geoinformatics. Previously he has worked in Gadjra Mada University's Department of Geography, Glasgow University's Topographic Science Section as an Assistant Lecturer, and at Fairey Surveys.

Du Daosheng is a staffmember of the National Key Laboratory at Wuhan Technical University of Surveying and Mapping, holding the rank of professor. Prior to that he was Head of the Department of Cartography at the same institute. From 1982 - 1984 he worked in the UK NERC's Experimental Cartography Unit.

Corné van Elzakker is a staffmember of ITC's Department of Geoinformatics. Prior to that he was a lecturer in the Cartography Section of the Department of Geography of Utrecht State University.

# Communication of Uncertainty in Spatial Data to Policy Makers

Morwenna Spear<sup>1</sup>, Jane Hall<sup>2</sup> and Richard Wadsworth<sup>2</sup>

---

**Abstract.** - There is increasing interest in the spatial aspects of many policy decisions. To be valid decision-making tools it is important that maps not only convey what is known, but also where the uncertainties lie. This paper reports the results of tests conducted to assess the ability of various methods of error representation to accurately assist interpretation of maps. The work uses the predicted surface and associated error surface (both generated by a kriging function) to test the ability of prospective and current users to interpret the data presented, and to draw simple logical conclusions about it. The ultimate aim is to study the effect of user perception of accuracy on the real-life decisions made using interpolated spatial data.

---

## INTRODUCTION

Maps are an effective method for communicating the spatial aspect of scientific data to policy makers. The current high priority on environmental issues means that maps may be required to show levels of pollution and their effects on different ecosystems to policy makers. Creation of these maps usually requires the use of models and interpolation techniques. This has been made easier as Geographical Information Systems (GIS) have evolved to handle the large quantities of data and perform elaborate calculations quickly. Unfortunately these developments have an associated cost, in that the new data processing methods may introduce and magnify uncertainties and errors in ways which previously have not required such consideration. Conversely, the enhanced data handling ability equips us to study these errors and uncertainties more easily and we therefore should begin to communicate this information to map users.

The analysis of sources of errors and their propagation has been an area of increasing interest since the mid 1980s (Heuvelink *et al.* 1989). Many researchers have investigated the topic and used various techniques (for example, see Veregin 1994). However, there have been fewer reported investigations into methods for displaying values of error and uncertainty. Maps may sometimes be presented with an estimated error, for example, digital elevation models published by the United States Geological Service are supplied

---

<sup>1</sup> Research Assistant, TRADA, High Wycombe, Bucks. United Kingdom

<sup>2</sup> Higher Scientific Officer, Institute of Terrestrial Ecology, Huntingdon, Cambs. United Kingdom

with an indication of their root-mean-squared-error. Any report of error is welcome, but as Lee *et al.* (1992) point out the magnitude of the error will vary over the area and a different approach may be more appropriate. In presenting interpolated data, associated error estimates are sometimes presented in map form (Webster *et al.* 1991). This allows the map user to be aware which regions within the area are likely to be prone to high errors, and to make a more informed decision in the light of this extra knowledge.

Work at the Environmental Information Centre of the Institute of Terrestrial Ecology includes presenting maps of environmental data to policy makers in Government departments. Typically these maps show levels of pollutants deposited generated by interpolating point data from across the UK. They may then be combined with suitable environmental receptor maps. The resulting maps may then be used to identify areas where the threshold of a pollutant is in excess of the ecosystem's ability to buffer it. This research has considered methods for presenting the uncertainty in interpolated data to decision makers. Several methods of communicating uncertainty in raster data sets have been developed and tested for their ability to convey understanding of error. Ultimately we hope to assess the sources and compounding factors for all errors associated with our data.

## **INVESTIGATION METHODS**

The emerging study area of error visualisation draws on knowledge from a wide range of disciplines: geostatistics, computing, cartography, psychology. In the development of new presentation techniques our aim was to inter-relate the knowledge and understanding of the various disciplines to ensure the solution found was reasonable.

### **Investigation of Interpolated Data Errors**

To gain a fuller understanding of the nature of the errors present in interpolated data we generated several known data sets, resembling interpolated deposition maps. These allowed better understanding of the errors which interpolation introduced. From these hypothetical maps we sampled  $z$ -values using various sampling schemes. Kriging was chosen for interpolation of these sampled data as it gave an accessible measure of uncertainty. Using this technique a number of predicted maps and predicted error values were generated. We were able to compare the estimates with the real values, which allowed us to verify that the errors predicted with the interpolation were appropriate. We found that the predicted error values were approximately equivalent to one standard error; in that sixty-eight percent of actual errors were smaller than the predicted error. This implied that the error values could be used to define a confidence

interval. This assumption was confirmed by finding that an interval twice the predicted error included 95% of actual errors. Using the predicted error values, 68% confidence intervals were calculated for each predicted map. It should be noted that further study will address the correct use of confidence intervals for modelled error. This paper is concerned with the validity of the visualisation methods.

### **Presentation of Estimated Uncertainty**

Past tests have shown that layer tint maps are the most successful method for visualisation of a landscape (Phillips *et al.* 1975, Phillips 1984). The colour of the layers reminds the map user of the value between the contour lines when scanning a large area of the map. The limitations when visualising a surface imply that the inclusion of uncertainty as an extra variable will pose greater problems for our visual memory. Layer tint maps were used as they are the least demanding method of display, as well as being the simplest way of handling raster data.

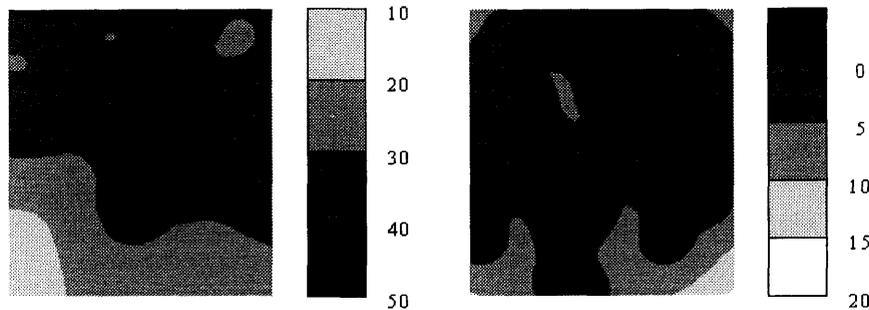
The simplest way of presenting the estimated uncertainty is as a separate map. However the definition of a confidence interval enables us to communicate the errors present in an entirely different manner. The use of raster data sets allows the uncertainty and confidence interval to be calculated on a cell-by-cell basis. The confidence interval can be presented in three separate maps (being the predicted map and the upper and lower limits of the interval - see figure 2), or the three maps can be combined into one image (figure 3). Either form of presentation should be accompanied by a statement of the level of confidence presented. Using the three maps separately, map users can be confident in their extraction of information. For example, the map user may check that they have selected the correct end of the confidence interval for their current use. Where the maps have been combined, the process gives rise to hatched areas where the classes from each of the three constituent maps overlap. The hatchings give an impression of the uncertainty of the information being presented, marking the extent to which the class boundary may differ in reality. It is possible to visualise the effect of the uncertainty in all three dimensions, without the need to cross refer between maps.

### **SURVEY METHODOLOGY**

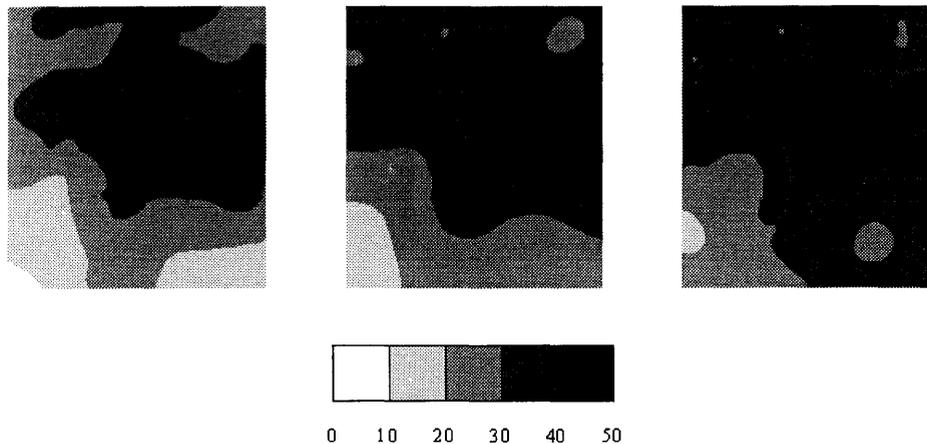
Having developed several visualisation methods the next phase of the investigation was to conduct a survey. This was designed to assess the level of information which could be conveyed about uncertainty in interpolated maps and test a representative selection of the visualisation techniques developed. The composition of tests which have been conducted in the past (Phillips *et al.* 1975,

Phillips 1982, Potash *et al.* 1978) were considered, and have been modified to reflect the requirements of presenting error to decision makers. We wished to check that information from the predicted map was still accessible, but more importantly to assess the ability of candidates to determine the effects of uncertainty on that prediction.

The three techniques selected to be investigated in our survey were presentation of the error on a map beside the predicted map (figure 1); presentation of a confidence interval as three separate maps (figure 2) and presentation of a confidence interval as one combined map (figure 3)



**Figure 1. - Presentation of interpolated prediction (left) and the predicted error (right).**

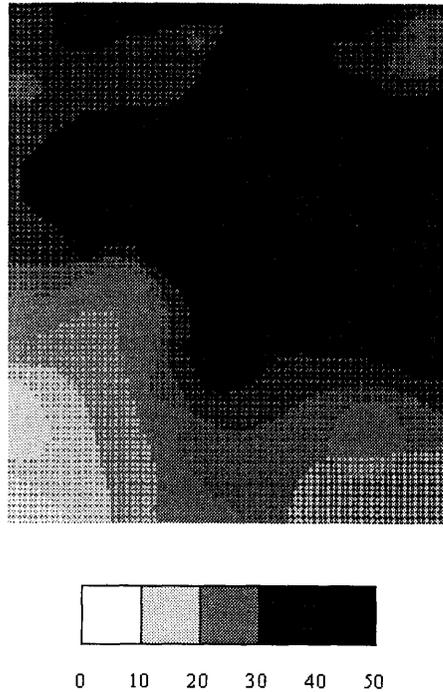


**Figure 2. - Presentation of interpolated prediction (centre) with the confidence interval limits; lower (left) and upper (right).**

### Structure of Survey

The survey consisted of two sections: assessment of users' skill and assessment of users' preference.

The first section assessed the candidate's ability to understand maps presented



**Figure 3. - Presentation of the confidence interval maps as a combined image.**

using the three techniques listed above. Some maps may be used to locate areas in which, for example, pollutant deposition values exceed an acceptable limit. One set of questions addressed this issue, asking the candidates to mark the areas in which the values might exceed a given level (when the error was taken into account). A second type of question asked the candidates to mark the highest and lowest points on the map, to assess whether they could understand the key and layer tint used. Both question types were repeated for all three methods, as well as for a control question in which no error was presented. A further type of question asked the candidates to identify areas where additional data samples were needed to reduce the resultant error.

Early in the first section, where the methods had been presented but not explained, the candidates were asked for their opinion of the methods. This was to gain an impression of their immediate reaction, simulating a situation where a new method of presentation was encountered in a journal or book without detailed explanation. The opportunity to give a more detailed response on the three methods was given at the end of the test after the candidate had had further experience with all three methods.

The second section contained questions which allowed the candidates to select the method with which they “felt most comfortable” to answer questions similar to those described above. Here, all three methods were given, on the same page, using maps of equal size. The candidates were not limited to the use of only one method but were asked to tick each map which they used. These

questions were used to assess the method the candidates preferred, and hence the method most likely to be popular in a real situation. The information was also used to check that the most popular method was actually the method which was used correctly.

Candidates were selected to represent as wide a selection of potential map users as possible. In all, 24 science and administrative staff participated, representing a range of mapping expertise and seniority. The level of geographical qualification held by each candidate varied from none to postgraduate. The frequency of map use varied from never to daily. The candidates were told that completing the questionnaire would take approximately 30 minutes, but that if they felt they required longer, or needed to finish sooner in order to meet their other work commitments, they could do so.

## **ANALYSIS OF RESULTS**

The candidates' responses have been used to investigate many aspects of the presentation techniques. Most important is the comparison between the methods using error maps with methods showing confidence intervals. The comparison between the two confidence interval presentations chosen is also interesting, but we anticipate future developments in the presentation of confidence intervals will lead to altered preferences. The results outlined below should be a starting point for further investigations.

### **Time and Experience**

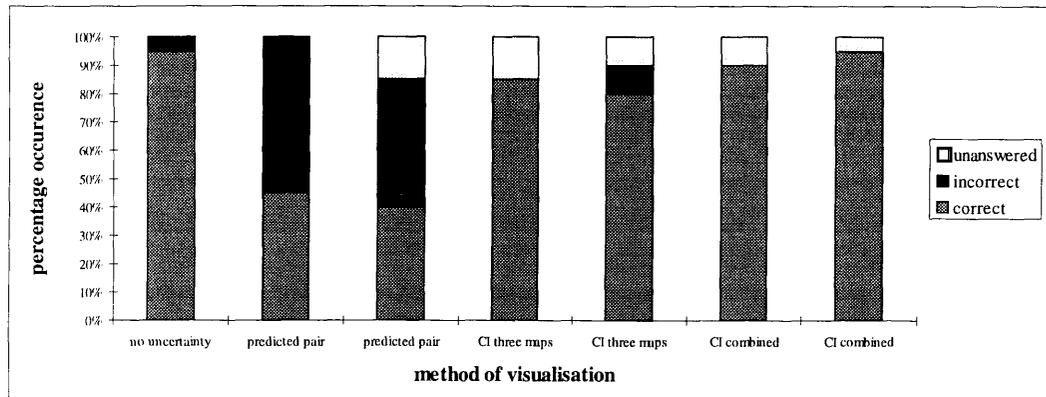
Before detailed analysis of the results, it was necessary to consider the effect of external factors on the candidates' responses. As was to be expected, the candidates' levels of geographical qualification and their frequency of map use were found to be related to their performance in the survey. However, the wide range of geographical skills which civil servants and policy makers possess, made it desirable to ensure that this range of skill was also present in the survey. The requirement for sufficient explanation of the occurrence and effects of uncertainty in maps cannot be understated as we must be sure that the topic is understood correctly.

It was also important to assess the effect of the time taken to complete the questionnaire. The quality of answer was shown to have no correlation with the amount of time spent in responding. So the lack of imposed time constraint appears to have had no biasing effect on the results of the survey.

### **Identification of Threshold Values**

The responses were scored by two methods. Initially they were assessed with

reference to the ideal response. A score was allocated to reflect the level to which the candidate had understood and implemented the idea. This indicated that the combined map gave the highest score for most candidates. This scoring method was quite arbitrary, so a second method was designed to be more quantitative. A regular grid of points was prepared on transparency, each point being marked as being above or below the threshold. The grid was overlaid on the response, and the identified points split into four categories: correctly marked, correctly unmarked, incorrectly marked and incorrectly unmarked. The counts were used to conduct chi-squared tests which identified the responses which were wrong. The percentage of candidates giving correct responses was used as an indicator of the success of the method. Occasionally a candidate had not answered the question or had marked the map in a way which was not suited for this type of analysis, so these responses have been treated as a separate category. The results for each presentation method are shown in figure 4.



**Figure 4. - Proportion of candidates identifying exceeded areas correctly. This question type was used once where no uncertainty was presented, then repeated twice for each of the three uncertainty display methods.**

The most visible trend was that the use of the predicted error map alongside the predicted map gave a significantly lower standard of result than where either form of confidence interval was used. Indeed, with the confidence interval presentations the scores were similar to the control example where no error was presented with the map. In situations where the map user requires a visual perception of the extremes which the uncertainty implies, the confidence interval is the more useful tool. The results also indicate that the combined map of the confidence interval was the most accurately used. This observation might be an artefact of the relatively small number of questions asked. Ideally further studies would be conducted to assess the most effective method amongst the possible confidence interval representations. Other methods for presenting the confidence interval are currently being developed.

## **Assessing Location For Extra Samples**

Where the candidates were requested to locate areas where extra samples were needed, the majority of candidates opted to use the predicted error map. There were however, two other methods which candidates justifiably chose to use. One was to refer to both the predicted error and the original map. This reduced the uncertainty in areas where it was high and verified the “topographical” features which the map seemed to predict. The other was to use the combined map to locate areas where the cells’ confidence interval spanned two or more class intervals. This was a very interesting option, as the tint layers would ideally be selected to reflect the classes in which the map user was interested. Areas which were borderline between two classes would be the areas in which the uncertainty would have the most influence over that decision. Clarifying the boundaries in these areas would be the most cost effective action.

## **Candidates’ Preferences**

Where candidates were allowed to choose between the three methods of presentation, the most popular method was the three separate confidence interval maps, but the use of the combined map came a close second. Some candidates opted to use both of these methods, and explained this choice, saying that the three separate maps were useful for ensuring that they selected the correct information and map. The combined map was additionally useful in spatially relating the three maps to the one area under consideration and identifying the areas on the original map.

The candidates’ opinions expressed in the free response questions often indicated quite clearly their preferences between methods. However a number of candidates made no comment. The trend for preferring the confidence interval was explained by the difficulty of superimposing the original two maps, and the simplicity of using the confidence interval. The reasons expressed for lack of conviction in the combined confidence interval were problems with the explanation, difficulties with the key and greyscale, and the presence of too much information. However a number of people expressed a preference for this method, as the shading makes it easier to visualise patterns.

## **CONCLUSIONS**

Currently the presentation of uncertainties and errors in maps is very limited. This situation has perpetuated many untrained users’ perceptions that maps are infallible. Where maps are subject to errors and uncertainties which might influence policy decisions, these now can and should be presented. The results of

the survey presented here show that it is possible to convey information on uncertainty and error, even to people who are unfamiliar with the data. The use of a confidence interval gives an enhanced understanding of the potential effect of the uncertainty on the reported values. There are many potential methods for presenting a confidence interval, the two which have been reported here have been investigated in depth, but others have also been considered by the authors. Other methods will doubtless be developed to satisfy specific requirements.

The presentation of the error as a map in its own right is useful in the planning stages of a sampling scheme for interpolation, or where a model needs to be improved. However, the use of the combined confidence interval map has proved more effective in targeting areas where the uncertainty has the greatest effect on the data currently under consideration. The use of a confidence interval can be employed in any raster data set, and the confidence interval can be defined according to the requirements of the users and the quality of the data. The use of a confidence interval in mapping is a relatively new concept which has the possibility of being a major advancement in understanding uncertainty within maps. As information on uncertainty in critical loads and deposition estimates becomes available the communication of uncertainty to policy makers will become a standard procedure.

## REFERENCES

- Heuvelink G.B.M., Burrough P.A. and Stein A. 1989 Propagation of errors in spatial modelling with GIS *International Journal of Geographical Information Systems* **3(4)** pp.303-322
- Lee J., Snyder P.K. and Fisher P.F. 1992 Modelling the effect of data errors on feature extraction from Digital Elevation Models *Photogrammetric Engineering and Remote Sensing* **58(10)** pp.1461-1467
- Phillips R.J., DeLucia A. and Skelton N. 1975 Some objective tests of the legibility of relief maps *Cartographic Journal* **12(1)** pp.39-46
- Phillips R.J. 1982 An experimental investigation of layer tints for relief maps in school atlases *Ergonomics* **25(12)** pp.1143-1154
- Potash L.M., Farrell J.P. and Jeffrey T. 1978 A technique for assessing map relief legibility *Cartographic Journal* **15(1)** pp.28-35
- Veregin H. 1994 Integration of simulation modelling and error propagation for the buffer operation in GIS *Photogrammetric Engineering and Remote Sensing* **60(4)** pp.427-435
- Webster R., Campbell G.W. and Irwin J.G. 1991 Spatial analysis and mapping the annual mean concentrations of acidity and major ions in precipitation over the United Kingdom in 1986 *Environmental Monitoring and Assessment* **16(1)** pp.1-17

# Measuring the Performance of Algorithms for Generating Ground Slope

William H. Ryder<sup>1</sup> and Demetra E. Voyadgis<sup>2</sup>

---

**Abstract.** - Calculation of ground slope is a fundamental geographic information system (GIS) utility. While accurate slope prediction is acknowledged to be a fundamental capability, reality is that users normally apply the slope utility in their GIS without considering accuracy of the product. As the Army's center of expertise for digital topographic data, the Digital Concepts and Analysis Center at the U.S. Army Topographic Engineering Center (TEC) has conducted a study of the accuracy of several slope generation algorithms. This study was designed to field test existing slope generation algorithms, to identify one or more that performed satisfactorily, and to recommend incorporation of those into Army systems. Several Army source algorithms as well as algorithms embedded in commercial off the shelf (COTS) software packages were tested. Rapid field collection of point slope data, a key aspect of this study, was made possible through Global Positioning System (GPS) and laser range finding technology. To assess slope algorithm accuracy, field slope measurements were compared with computer generated slope values.

---

## INTRODUCTION

Detailed mapping of the Earth's surface configuration (slope) has always been time intensive. Traditional slope compilation methods are very labor intensive, requiring analysis of contour line information by a human interpreter. Compilation time can be dramatically reduced by using a computer to generate slope from digital elevation data. Key ingredients in this process are elevation data with adequate resolution and an algorithm that performs satisfactorily. This study was commissioned to examine algorithms for slope generation from elevation data.

Automated slope generation has been a topic of research for many years. Efforts by the Defense Mapping Agency (DMA) and the U.S. Army Topographic Engineering Center (TEC) during the 1970's were abandoned due to the lack of accurate Digital Terrain Elevation Data (DTED). Since then, source data has

---

<sup>1</sup> Physical Scientist, U.S. Army Topographic Engineering Center, Alexandria, VA.

<sup>2</sup> Geographer, U.S. Army Topographic Engineering Center, Alexandria, VA.

improved and new algorithms have been written. Problems also existed because users wanted a product that resembled a traditional, manually compiled slope overlay and slope generated from gridded elevation data was blocky and disjointed. This is less of a problem today because slope is mainly used in computer models and seldom as a stand-alone product. With higher spatial resolution elevation data and faster computers, automated slope generation is a reality. The question remains, given identical elevation data inputs, which algorithm creates the best representation of the Earth's surface configuration?

## APPROACH

The approach used by this study was to review literature on algorithms used currently, and to select candidates based on the availability and acceptance by the users. Three grid-based algorithms were examined. Also examined were two manually compiled slope overlays. (See Table 1).

**Table 1.--Slope Generation Methods.**

System/Author	Method used
ARC/INFO	Uses the average maximum technique. Calculates slope from the third order finite difference method as described by Horn, 1969. Doubles the weight of the nearest neighbors in the cardinal directions.
ERDAS IMAGINE	Similar to ARC/INFO, but weighs all neighbors equally.
U.S. Army Waterways Experiment Station (WES)	Calculates slope from normal vectors and the second order finite difference method.
1:24,000 Slope Overlay	Manually compiled from the contours of a 1:24,000 USGS Quadrangle.
Tactical Terrain Analysis Data Base Slope Overlay	Manually compiled by DMA from the contours of a 1:50,000 Topographic Line Map.

The algorithms were tested using a high spatial resolution elevation data set. Elevation postings for this data set were collected at five meter spacings. Besides generating slope using the five meter data, spacings were thinned to 30 and 100 meters. This was done to approximate the elevation post spacing of DTED Levels 2 and 1 respectively.

The slope values that resulted from the different algorithms and resolutions were categorized according to the ranges used on a DMA Tactical Terrain Analysis Data Base (TTADB) (See Table 2). An TTADB is a collection of six overlays (obstacles,

soils, slope, surface drainage, transportation and vegetation) used by the military for tactical planning. The categorized results were compared statistically against field data to calculate the overall accuracy.

**Table 2.--TTADB Slope Categories**

<b>Slope category</b>	<b>Range (percent slope)</b>
A	$0 \leq 3$
B	$>3 \leq 10$
C	$>10 \leq 20$
D	$>20 \leq 30$
E	$>30 \leq 45$
F	$>45$

## **DATA**

### **Data Used**

To meet study objectives, data used had to be accurate and have high spatial resolution. Spatial resolution had to be at least equal to that of DTED Level 2 ( 30 meters). Data that met these requirements were available over the Yakima Training Center, Washington. An elevation matrix was created on a digital stereo photogrammetric workstation using GPS controlled 9" by 9" imagery.

The Yakima Training Center was chosen for this study because of the variety of slope categories found there. Additionally, the road network throughout the area allowed for easier access to site locations. To conserve computer resources, a four kilometer by four kilometer area was used as the test area. This area contained the required variety of terrain/slope categories.

### **Collection of Field Data**

The procedure for collection of data consisted of two parts, test point selection and slope measurement in the field.

#### **Test Point Selection**

Field objectives of this study required that each point visited in the field meet the following conditions: points had to be separated by at least 125 meters (this was to ensure that the longest distance measured, 100 meters, did not fall into area covered by other points), and the 125-meter radius circle surrounding the point had to be totally contained within a slope category, not crossing over into another category. To meet these conditions, approximately 50 points from each slope category were preselected on the 1:24,000 slope overlay. The actual number of points visited was lower because of firing range restrictions at Yakima.

Once in the field, personnel navigated to the preselected coordinates using a Precision Lightweight Global Positioning System Receiver (PLGR). When using the military key, the PLGR is accurate to within ten meters. This amount of accuracy is sufficient for measuring slope values at thirty and one hundred meter post spacings, but higher accuracies were needed to measure slope at five meter spacings. To achieve higher accuracies, differential GPS was used. After reaching a point using the PLGR, readings were taken using the differential GPS unit. This unit took position readings every five seconds until sixty positions were recorded. These positions were later differentially corrected by comparing them with readings taken by a GPS base station set up over a known point. The average of the corrected readings provided a coordinate that was accurate to approximately two and one-half meters.

### **Field Measurement of Slope Values**

Slope was measured in the field using laser ranging equipment. Once a point was found using the PLGRs, and its actual position was recorded with the differential GPS, one analyst remained on the point while the other analyst paced-off a distance of five meters in the direction of the steepest slope. The exact distance was measured using the range gun, and the analyst's position was adjusted accordingly. When using the range guns, care was taken to ensure that the measurement point on the analyst was at the same height as the gun itself. Maximum slope, when not readily apparent, was determined by measuring in several directions. The laser ranging equipment is ideal for this work, instantly giving slope percent and horizontal distance as well as azimuth. After recording the readings, the analyst moved to thirty meters from the start point, again making sure to be on the maximum slope, and the measurements were taken. This procedure was repeated at 100 meters from the start point. Maximum slope was determined for each distance. The maximum was not always in the same direction for all distances.

### **MANUALLY GENERATED SLOPE METHODOLOGY**

The 1:24,000 slope overlay was manually compiled in-house, using the contours from a United States Geological Survey 1:24,000 topographic map and a slope wedge. The slope wedge uses a series of circles whose sizes correspond to the slope categories. The category is determined by the largest circle that can fit between two contours. As the distance between the two contours increases or decreases, the slope category changes. A minimum polygon size of 250 meters by 250 meters (ground distance) was used in accordance with DMA specifications. This is the same method that DMA used to compile the TTADB from the contours of a 1:50,000 Topographic Line Map.

## COMPUTER GENERATED SLOPE METHODOLOGY

Three methods were used for generating the slope thematic maps: an ARC/INFO method, a WES method, and an ERDAS IMAGINE method.

### The ARC/INFO Method

A matrix of the elevations points was output in ASCII format. The ARC *generate* command was used to generate a point coverage, which was brought into the GRID module through the *pointgrid* function. Next, the field locations were downloaded from the differential GPS in ASCII format. Again, the *generate* command was used to create a point coverage, the *pointgrid* command used to create a grid coverage.

The original elevation data set, at 5-meter spacing, was thinned to 30 meters and 100 meters using the *resample* command. For each of these three data sets, the following procedures were performed:

- define the projection
- calculate the slope
- classify the data
- extract the slope category values for the field locations.

The *projectdefine* command was used to set the projection, datum, and units of measure to match those used during data extraction. (This was also performed on the field location data set.) The slope was calculated using the *slope* command. A *remap table* of values, using the TTADB slope categories, was created and used with the *reclass* function to produce the thematic map. Finally the *combine* command was used to extract the slope category values. The *combine* command combines two or more coverages and outputs a coverage of unique values. The database table associated with the resulting coverage contains the values from each of the original coverages. Therefore, when the reclassified slope grid and the field location grid are input to the *combine* function, the resulting grid coverage contains only those grid cells that are common to both input grids. The database table for the resulting grid contains the slope category value and the field location number. These numbers were then used in the statistical comparisons.

### The WES Method

The WES algorithm is a stand-alone utility that generates slope values directly from a matrix of elevation points, and returns the results as a matrix in ASCII format. Therefore, a software suite was needed to classify the slope values and to extract the values for the field locations. ARC/INFO was used for these purposes. Using the ARC/INFO commands stated above, the following procedures were performed on the algorithm at all three resolutions (5, 30, 100 meters):

- define the projection
- classify the data
- extract the slope category values for the field locations.

These slope category values for the field locations were then used in the statistical comparisons.

### **The ERDAS IMAGINE Method**

The original 5-meter point file generated in ARC/INFO from the ASCII output, was imported into IMAGINE. An image file was created at a 5-meter cell size and the *resample* function was used to thin the data set to 30 and 100-meter resolutions. Also imported was the field location point file generated in ARC/INFO. The same procedures as in ARC/INFO were used for the elevation data sets at all three resolutions.

Using the *image info* utility, the projection, datum, and units of measure were defined for the elevation data sets (as well as the field location data set) in accordance with those parameters used during the extraction process. The *raster recode* function was used to reclassify the data into TTADB slope categories. Finally, a spatial model was created with the *model* utility using the *conditional* function. Using the slope data and field points as input, the *conditional* function outputs the pixel value of the first input grid for every pixel value existing in the second input grid. Therefore, a slope value was output for every field location. The *pixel to table* function was used on the resulting image file to create an ASCII file that contained the x, y, z values for each pixel. A z value of zero indicated that the pixel was not a field location point. To extract the field location points, the UNIX *awk* command was used to output the x, y, z values for each pixel with a non-zero z value. These values were then used in the statistical comparisons.

## **STATISTICAL ANALYSIS**

The slope values generated by the algorithms at the selected points were compared with the values obtained in the field by using a confusion matrix. The generated values were matched against the known (field measurement) values, allowing an assessment of slope category accuracy and overall accuracy. These results were used to judge the performance of the algorithm.

**Table 3.--Sample confusion matrix.**

		Generated Values							
K n o w n  V a l u e s		A	B	C	D	E	F	Row Total	
	A	<b>32</b>	11						43
	B	4	<b>38</b>	6					48
	C	1	15	<b>28</b>	4	1			49
	D	1	1	11	<b>15</b>	5			33
	E			1	11	<b>14</b>			26
	F			2					2
	Column Total	38	65	48	30	20	0		201

The bold numbers in the sample above are the number of points correctly classified. This means that at a given point, the generated value matched the field measurements. In the sample, 32 points had the correct "A" value. Eleven points were classified as "B" values by the algorithm while they were actually "A" values according to field measurements.

To calculate the percent correct in any given category, divide the total correctly classified for the category by the row total (i.e.,  $32/43 = 0.744$  or 74.4% correctly classified as "A" slope). To calculate the overall accuracy, the total correctly classified for the matrix (sum of the diagonals) is divided by the total of all the values in the matrix (i.e.,  $(32 + 38 + 28 + 15 + 14)/201 = 0.632$  or 63.2% of the values were correctly classified by this algorithm).

## RESULTS

The prediction accuracies of the methods tested are shown in Table 4.

## DISCUSSION

The slope generation algorithm embedded in ARC/INFO outperformed all others at 5 and 100-meter resolutions. At 30-meter resolution, the ARC/INFO algorithm also performed well though was slightly less accurate than the WES algorithm (74.6% versus 76.1%). The WES algorithm was a consistently good performer, but failed to classify any "B" slopes at 5 meters, indicating a possible flaw in the software. Further testing must be done before any recommendation can be made. All of the algorithms tested, except the WES tended to underestimate slope. The WES algorithm overestimated slope, especially in the "B" category. The manually

**Table 4.--Results of testing**

ELEVATION RESOLUTION	METHOD	CATEGORY ACCURACY (%)						OVERALL ACCURACY (%)
		A	B	C	D	E	F	
5 Meters	ARC/INFO	74.4	79.2	57.1	45.5	53.8	0.0	63.2
	IMAGINE	65.8	63.0	41.2	38.2	48.0	0.0	50.8
	WES	75.0	0.0	83.3	45.7	69.2	0.0	51.7
	1:24,000	81.6	46.7	48.1	37.5	13.0	60.0	48.3
	TTADB	84.6	38.6	32.1	37.8	40.0	0.0	44.8
30 Meters	ARC/INFO	86.7	85.2	68.8	56.2	63.6	--	74.6
	IMAGINE	93.6	82.0	59.6	48.4	50.0	--	70.3
	WES	79.5	84.9	71.4	69.7	68.2	--	76.1
	1:24,000	90.9	69.8	58.6	45.5	12.0	--	60.1
	TTADB	88.6	51.1	41.8	51.4	54.2	--	57.1
100 Meters	ARC/INFO	93.5	83.9	70.7	46.4	23.1	--	73.1
	IMAGINE	93.2	75.4	52.5	41.4	7.7	--	63.4
	WES	65.9	76.9	77.8	66.7	50.0	--	71.6
	1:24,000	93.3	64.4	56.9	36.4	13.3	--	60.1
	TTADB	88.9	44.7	39.7	48.5	66.7	--	55.2

compiled overlays also overestimated slope.

Overall, accuracy of the algorithms at 30 and 100-meter resolutions was better than the accuracy achieved at 5 meters, when compared with the field collected data. A possible reason for this lies in the accuracy of the coordinates obtained using differential GPS. Although the method produced ground coordinates with an accuracy of +/- 2.5 meters, this may not have been sufficient accuracy for comparisons to the 5-meter grid. Due to the 2.5-meter uncertainty in the location of the measuring point, exact correlation with the automated slope prediction grid was not possible. This discrepancy may be responsible for misclassification of some points.

As shown in Table 4, the manually compiled overlays at 1:24,000 and 1:50,000 scale (TTADB) were less accurate slope predictors than all of the algorithms. As expected, the larger scale overlay (1:24,000) did predict better than the smaller scale overlay (1:50,000). Part of the reason for the observed lower accuracy of these overlays is the minimum compilation size of each polygon. For this study, the

smallest area retained during compilation was 250 by 250 meters on the ground. This minimum size affects the prediction accuracy though the overriding factor is that manual compilation is clearly less accurate than automated approaches.

All algorithms tested to this point are grid-based. The need to examine algorithms utilizing triangulated irregular networks (TINs) to compute slope still exists. Additionally, TINs using geomorphic data (such as drainage networks) must be tested to see if this additional data improves the accuracy of the slope predictions.

## CONCLUSIONS

1. Use of grid-based slope algorithms alone will produce point slope predictions that match the TTADB slope categories approximately 75% of the time.
2. The WES Maximum slope and the ARC/INFO algorithms performed similarly except for the "B" slope anomaly for the WES Maximum slope algorithm noted earlier in the discussion.
3. The ERDAS IMAGINE algorithm produced slightly lower prediction accuracies than the WES Maximum slope and the ARC/INFO algorithms.
4. All of the automated slope prediction algorithms outperformed the manual slope compilation approach.
5. Use of TIN-based algorithms with and without geomorphic data may increase prediction accuracy.

## ACKNOWLEDGMENTS

The authors of this paper would like to thank Louis Fatale and James Ackeret for their excellent field support, and Patrick Nguyen for his system support. We also greatly appreciate the leadership and support of Mr. Jeffrey Messmore.

## REFERENCES

- Burrough, P.A. 1986. Principles of Geographical Information Systems for Land Resources Assessment, New York, NY: Oxford University Press. pp. 39-56.
- Environmental Systems Research Institute. 1994. Help documentation for ARC/INFO 7.0.2, Redlands, CA.
- ERDAS Inc. 1995. ERDAS Field Guide, Atlanta, GA. pp. 322-324.
- Horn, B.K.P. 1969. Hill shading and the reflectance map, Proceedings of the I.E.E.E. pp. 14-47.
- Skidmore, A.K. 1989. A comparison of techniques for calculating gradient and aspect from a gridded digital elevation model, International Journal of Geographical Information Systems, Vol. 3, No. 4. pp. 323-334.

# Experimental Development of a Model of Vector Data Uncertainty

G.J. Hunter<sup>1</sup>, B. Höck<sup>2</sup>, M. Robey<sup>3</sup> and M.F. Goodchild<sup>4</sup>

---

**Abstract.**—In a recent paper by Hunter and Goodchild (1996), a model of vector data uncertainty was proposed and its conceptual design and likely manner of implementation were discussed. The model allows for probabilistic distortion of point, line and polygon features through the creation of separate, random horizontal positional error fields in the  $x$  and  $y$  directions. These are overlaid with the vector data so as to apply coordinate shifts to all nodes and vertices to establish new versions of the original data set. By studying the variation in the family of outputs derived from the distorted input data, an assessment may be made of the uncertainty associated with the final product. This paper is a continuation of that initial work and discusses the experimental development undertaken thus far to implement the model in practice.

---

## INTRODUCTION

In a recent paper by Hunter and Goodchild (1996), a model of vector data uncertainty was proposed and its conceptual design and likely manner of implementation and application were discussed. This paper is a continuation of that initial work and discusses the experimental development since undertaken to implement the model in practice. In the context of this research, we suggest there is a clear distinction to be made between 'error' and 'uncertainty', since the former implies that some degree of knowledge has been attained about differences between actual results or observations and the truth to which they pertain. On the other hand, 'uncertainty' conveys the fact that it is the lack of such knowledge which is responsible for hesitancy in accepting those same results or observations without caution, and the term 'error' is often used when it would be more appropriate to use 'uncertainty'.

The model that has been developed can be defined as a stochastic process capable of generating a population of distorted versions of the same reality (such as a map), with each version being a sample from the same population. The traditional Gaussian model (where the mean of the population estimates the true value and the standard deviation is a measure of variation in the observations) is one attempt at describing error, but it is global in nature and says nothing about local variations or the processes by which error may have accumulated.

The model applied here is viewed as an advance on that approach since it has the ability to show spatial variation in uncertainty, and the capability to include in its realizations, the probable effects of error propagation resulting from the

---

<sup>1</sup> Assistant Professor, Department of Geomatics, The University of Melbourne, Victoria, Australia

<sup>2</sup> Scientist, Forest Research Institute, Rotorua, New Zealand

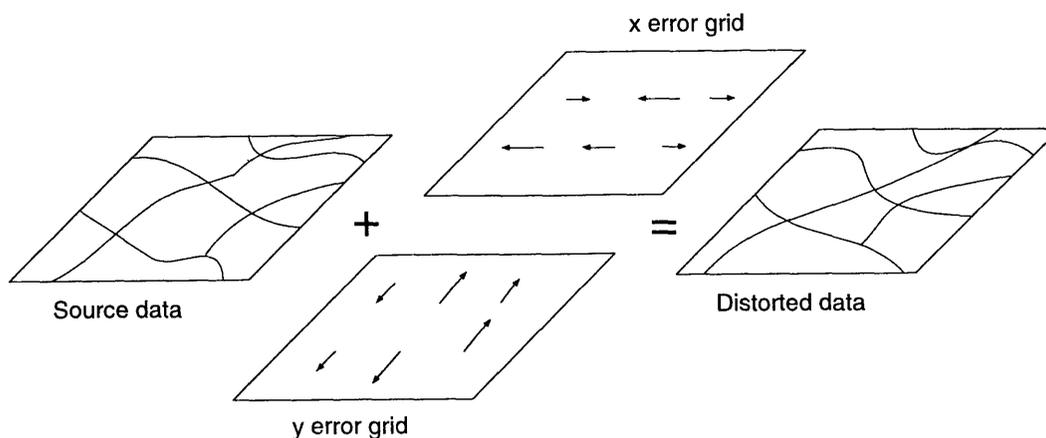
<sup>3</sup> PhD student, Department of Geomatics, The University of Melbourne, Victoria, Australia

<sup>4</sup> Director, National Center for Geographic Information & Analysis, University of California, Santa Barbara, CA

subsequent algorithms and processes that have been applied to the data. This latter point is particularly important, since most software vendors do not divulge the algorithms used in their packages for commercial reasons—which prevents formal mathematical error propagation analysis from being undertaken. By studying different versions of the products created by the model, it is possible to see how differences in output are affected by variations in input. Elsewhere in this proceedings a grid-cell version of the model (which is at a more advanced stage of development), has been described and applied to assess the uncertainty associated with landslide susceptibility indices derived from a simple slope stability model (Murillo and Hunter, 1996).

The vector data model involves the creation of two independent, normally distributed, random error grids in the  $x$  and  $y$  directions. When combined, these grids provide the two components of a set of simulated positional error vectors regularly distributed throughout the region of the data set to be perturbed (Figure 1). The assumptions made by the authors are firstly that the error has a circular normal distribution, and secondly that its  $x$  and  $y$  components are independent of each other. The grids are generated with a mean and standard deviation equal to the data producer's estimate for positional error in the data set to be perturbed (a prerequisite for application of the model). These error estimates, for example, might come from the residuals at control points reported during digitiser setup, or from existing data quality statements such as those that now accompany many spatial data sets.

By overlaying the two grids with the vector data to be distorted,  $x$  and  $y$  positional shifts can be applied to the coordinates of each node and vertex in the data set to create a new, but equally probable, version of it. Thus, the probabilistic coordinates of a point are considered to be  $(x + \text{error}, y + \text{error})$ . With the distorted version of the data, the user then applies the same set of procedures as required previously to create the final product, and by using a number of these distorted data sets the uncertainty residing in the end product is capable of being assessed. Alternatively, several different data sets may be separately distorted (each on the basis of its own error estimate) prior to being



**Figure 1.**—The model of vector data uncertainty uses normally distributed, random error grids in the  $x$  and  $y$  directions to produce a distorted version of the original data.

combined to assess final output uncertainty. While the model requires an initial error estimate for creation of the two distortion grids, it is the resultant uncertainty arising from the use of perturbed data due to simulation which is under investigation (in conjunction with the spatial operations that are subsequently applied)—hence its label as an ‘uncertainty’ model.

## DEVELOPMENT OF THE MODEL

### Choosing the Error Grid Spacing

As discussed in Hunter and Goodchild (1996), the first step required to implement the model is to determine an appropriate error grid spacing. If it is too large, the nodes and vertices of small features in the source data will receive similar-sized shifts in  $x$  and  $y$  during perturbation and the process will not be random, giving rise to unwanted local autocorrelation between shifts (Figure 2). Conversely, if it is too small then processing time can increase dramatically as additional grid points are needlessly processed.

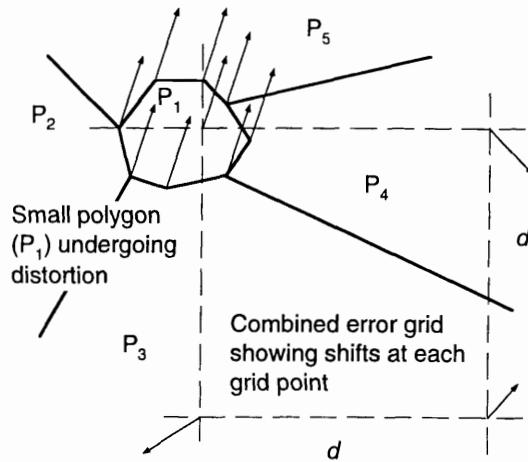


Figure 2.—If the error grid spacing ( $d$ ) is too large, then unwanted local autocorrelation between shifts may occur in small features.

The authors suggest an appropriate spacing be selected from one of the following options: (1) the standard deviation of the horizontal positional error for the source data; (2) a distance equal to 0.5mm at source map scale where the data has been digitized; or (3) a threshold value smaller than the user would care to consider given the nature of the data to be processed. Thus far, the authors have found the standard deviation (as supplied by the data producer) to be a useful value to apply, since it tends to be smaller than (2) and therefore more conservative, and of similar magnitude to practical estimates of (3).

### Generating the Initial Error Grids

Using the development diagram on the following page as a guide for the remaining discussion (Figure 3), the next step is to generate the  $x$  and  $y$  error grids. The Arc/Info and Arc Grid software modules were used as the development platform and the commands referred to in the paper apply to these packages. To ensure the grids completely cover the extent of the source data, the dimensions of the grids were pre-determined by setting a window equal to the data set's dimensions (SETWINDOW command), and the cell size equivalent to the chosen grid spacing (SETCELL command). Two grids (named accordingly for the  $x$  and  $y$  directions) were created automatically using these parameters and then populated with randomly placed, normally distributed values having a mean of zero and a standard deviation defined by the data producer (NORMAL

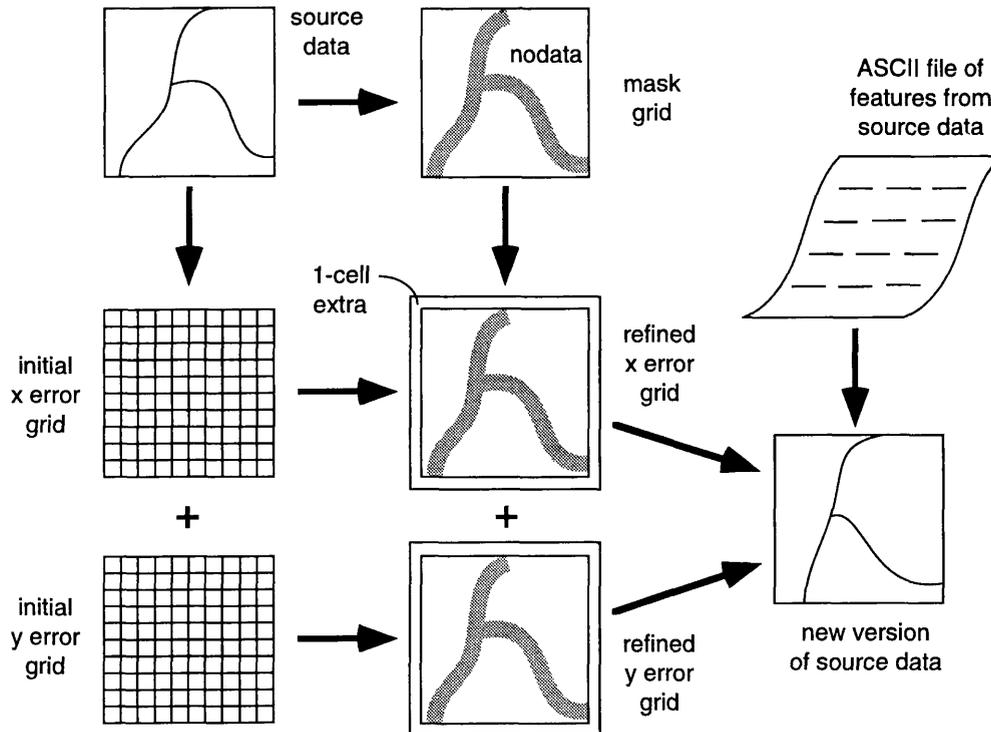


Figure 3.—Development diagram for the uncertainty model.

command). It should be noted here that selection of the standard deviation for the grid spacing does not affect the random population of the grids. The two grids are only temporary and will require further refinement before being used to perturb the vector data.

### Reduction of Error Grid Size

As a means of minimizing processing time the number of cells in the error grids needs to be reduced, since unless the vector data set is extremely dense there will be many cells that are not required during operation. To achieve this, the original vector data are converted to grid format (POINTGRID or LINEGRID commands) to form a temporary masking grid which only contains 'live' cells that the source data either lie within or pass through. Note that polygons are also processed by LINEGRID since only boundary strings need be perturbed. The cell attribute that has been passed during rasterization is unimportant since the grid will only be used for masking purposes. All other (non-contributing) cells are deliberately given a NODATA value.

However, a problem arises here with using a masking grid that contains only single cells (representing point data) or 1-cell-wide strings (representing lines or polygons). As mentioned in Hunter and Goodchild (1996), there is the possibility that the magnitude and direction of adjacent x and y error grid shifts may cause them to overlap each other, resulting in possible loss of topological integrity in the original data when applied to it (such as the transposition of adjacent vertices

causing unwanted loops).

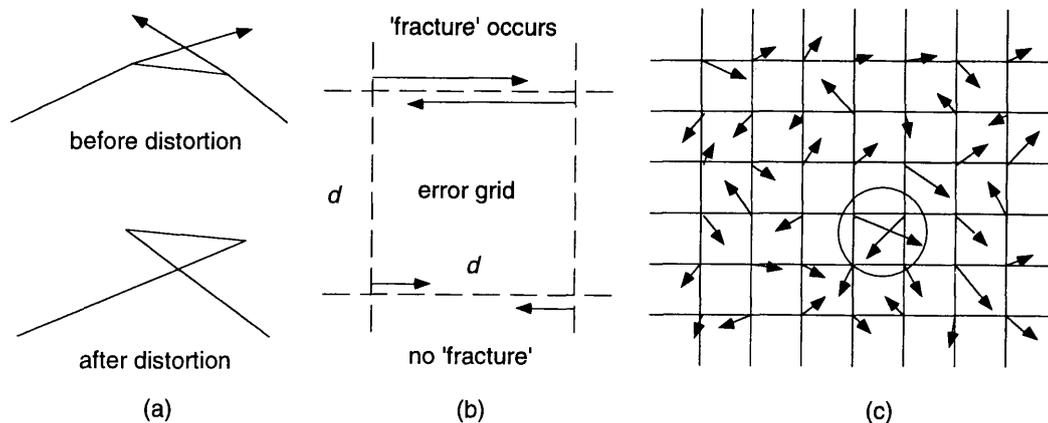
The solution requires filtering of any such ‘offending’ pairs of shifts which requires that the masking grid be extended for a distance of at least five standard deviations either side of the original data—given that with the normal distribution it is rare for a value to occur greater than this distance from the mean. This buffering of the masking grid will permit neighboring error grid shifts to be used in a subsequent filtering process.

To achieve this, the grid spacing is compared with the standard deviation of the data and the equivalent number of cells is calculated (remembering that the grid spacing will not always equal the standard deviation). For example, using a standard deviation of 20 m and a grid spacing of 30 m, the correct number of cells equals  $\lceil (100/30) + 1 \rceil = 4$  (with the extra cell and the modulus sign ensuring ‘rounding up’ of the answer). Using this value, a spread function is applied to the masking grid (EUCDISTANCE command). A new masking grid is created during the process and any (previously) NODATA cells affected by the operation are automatically returned to active status.

The masking grid is then overlaid with the initial error grids to provide reduced versions of the  $x$  and  $y$  grids that contain only shift values lying within the required regions. Finally, the error grids are expanded by the width of a cell on all sides (and given NODATA values) to support later processing (SELECTBOX command).

### Filtering the Error Grids

To preserve topological integrity in the source data upon distortion, some means of adjustment must be introduced to control the magnitude of positional shifts between neighbouring points in the  $x$  and  $y$  error grids. As previously mentioned, if this does not occur there is a likelihood that neighbouring points in the original data set may be transposed in position—causing unwanted ‘fractures’



**Figure 4.—In (a), uncontrolled shifts between neighbouring error grid points can cause unwanted ‘fractures’ or transposition of features. In (b), ‘fractures’ occur when the difference between neighbouring shifts is larger than their separation ( $d$ ). In (c), a ‘fracture’ is circled, requiring filtering on the basis of neighbouring shifts.**

in the feature structure (Figure 4a). Intuitively, these modifications should not occur and are considered unacceptable. Originally, the authors considered using spatial autocorrelation to constrain the  $x$  and  $y$  shifts, but this would have had the effect of unnecessarily altering all shifts in the error grids when only a relatively small number needed to be adjusted in practice. Accordingly, a localised filter was seen as a better solution.

Therefore, a routine was developed to test the difference between consecutive cells (in horizontal or row sequence for the  $x$  grid, and vertical or column sequence for the  $y$  grid) to determine whether the absolute value of the difference between them was greater than their separation distance (Figure 4b). If so, then a ‘fracture’ is possible at that location and a filter must be applied to average out the shift values on the basis of their neighbors. The procedure is iterative and proceeds until no ‘fractures’ exist in the error grids.

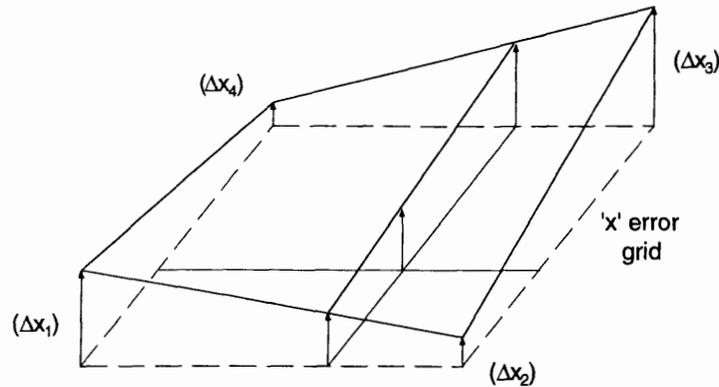
For example, given consecutive shifts  $\Delta x_0$ ,  $\Delta x_1$ ,  $\Delta x_2$  and  $\Delta x_3$  in any row of the  $x$  error grid (and testing the difference between the two middle shifts), there is potential for a ‘fracture’ to occur if  $\Delta x_1 - \Delta x_2 > d$  (where  $d$  is the grid spacing). The adjusted values of  $\Delta x_1$  and  $\Delta x_2$  are computed by equation (1):

$$\Delta x_1' = \frac{\Delta x_0 + \Delta x_1 + \Delta x_2}{3} \quad \text{and} \quad \Delta x_2' = \frac{\Delta x_1 + \Delta x_2 + \Delta x_3}{3} \quad (1)$$

The justification for the filtering is that although the shifts are normally distributed, the choice of that distribution is only an assumption and we should accept there may be faults in its selection as a theoretical model. For example, statisticians developed the human Intelligent Quotient (IQ) using a normal distribution with a mean of 100 and standard deviation of 20. Yet the distribution is deliberately truncated at zero to disallow negative values which intuitively should not occur—even though they will be predicted by the model. In a similar manner, the filtering of neighboring error shifts to avoid ‘fractures’ is simply a truncation of the distribution to cater for a minority of outlying values.

The filtering was implemented by developing a routine that moves progressively through each error grid and tests the difference between each pair of cell values (DOCELL command). If the condition for a ‘fracture’ is not met, the values are written to an update grid and the next pair is tested. If the condition is met, the filter is applied and new values are computed and passed to the update grid. In practice, it has been found that shifts of several standard deviations in size can cause several iterations of the process to be required. However, the procedure still converges fairly rapidly.

An early problem encountered was the situation where a large shift existed in a perimeter row or column of the error grid—requiring an extra cell value (which did not exist) to be input to the filter equation. It was found that without the placement of an extra row or column on each boundary of the grid, the absence of a cell caused the filter process to substitute a NODATA value for the cell to be adjusted. This was unacceptable and the placement of an extra perimeter cell around the error grid (with NODATA values) ensured that the original large shift was retained unaltered. This was considered preferable to it being discarded. While this implies that some nodes and vertices may receive shifts to new



**Figure 5.—The  $x$  and  $y$  shifts for a data point not coinciding with the error grid are calculated using bilinear interpolation based on the four surrounding grid values.**

positions outside the original extent of the data set, this decision remains consistent with the original concept of distorting their values to form alternative, but equally probable versions of the data. At the time of writing, the experimental development of the model has not proceeded beyond this stage and the following discussion remains conceptual.

### **Calculation of Data Point Shifts**

In the final step of the model's development, values in the error grids must be transferred to the data set being perturbed. Inevitably, it is expected that few if any nodes or vertices in the source data will coincide exactly with the error grid points, and a method is required for calculating  $x$  and  $y$  shifts based on the neighboring values in the grid. A simple bilinear interpolation procedure (Watson, 1992) is proposed in which the  $x$  and  $y$  shifts assigned to each point are calculated on the basis of the respective shifts of the four surrounding grid points (Figure 5).

An ASCII feature file containing the identifier and coordinates of each data point can be automatically derived (UNGENERATE command), and the four surrounding error shifts will be determined for each point and used to interpolate the correct values to be applied. The distorted coordinates are then written to a new file and when all records are processed, the file topology is rebuilt. Clearly, the need may arise for high performance computers to be employed for this task, particularly when many thousands of nodes and vertices may have to be perturbed, and current research into GIS and supercomputers (such as described in Armstrong, 1994) is being investigated by the authors.

### **CONCLUSIONS**

This paper has described the experimental development of an uncertainty model for vector data, which operates by taking an input data set of point, line or polygon features and then applying simulated positional error shifts in the  $x$  and  $y$  directions to calculate new coordinates for each node and vertex. In effect this produces a distorted, but equally probable, representation of the data set that can be used to create a family of alternative outputs, usually in map form.

Assessment of the variation in the outputs can be used to provide an estimate of the uncertainty residing in them, based on the error in the source data and its propagation through the subsequent algorithms and processes employed.

At this stage of the research, software has been written that automates the creation of the error grids and filters their values to avoid conditions that would cause loss of topological integrity in the original data. The software also masks the input data to ensure redundant values in the error grids are not unnecessarily processed. Optimisation of the process is still required to achieve faster operation. The next stage of the project involves development of an algorithm which enables each node and vertex in the input data to be allocated the correct  $x$  and  $y$  shift from the error grids, which yields the perturbed version of the source data. The final stage entails testing of the working model, refinement of its operation, and assessment of its strengths and limitations.

### ACKNOWLEDGMENTS

The authors wish to acknowledge the support given to this project by the New Zealand Forest Research Institute, the Australian Research Council, and the Australian Federal Department of Industry, Science & Technology (under the Bilateral Science & Technology Collaboration Scheme). The National Center for Geographic Information & Analysis is supported by the National Science Foundation, grant SBR 88-10917, and this research constitutes part of the Center's Research Initiative 7 on Visualisation of Spatial Data Quality.

### REFERENCES

- Armstrong, M.P. 1994. "GIS and High Performance Computing". *Proceedings of the GIS/LIS 94 Conference*, Phoenix, Arizona, pp. 4-13.
- Hunter, G.J and Goodchild, M.F. 1996. "A New Model for Handling Vector Data Uncertainty in GIS". *Urban and Regional Information Systems Journal*, vol. 8, no. 1, 12 pp.
- Murillo, M. and Hunter, G.J. 1996. "Modeling Slope Stability Uncertainty: A Case Study at the H.J. Andrews Long-Term Ecological Research Site, Oregon". *Proceedings of the 2nd International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Modeling*, Fort Collins, Colorado, 10 pp.
- Watson, D.F., 1992. *Contouring: A Guide to the Analysis and Display of Spatial Data* (Pergamon/Elsevier Science: New York).

# A User-friendly Tool for Error Modelling and Error Propagation in a GIS Environment

Frank Forier<sup>1</sup> and Frank Canters<sup>2</sup>

---

**Abstract.** - In spite of a growing concern for spatial database accuracy today most commercial GIS software products do not incorporate tools for the modelling of error in individual data layers and for the tracking of error when data layers are manipulated and combined in GIS-based spatial analysis. This paper reports on design considerations that have been taken into account in the development of a system-independent tool for error modelling and error propagation in raster-formatted categorical data. The tool aims at an optimal use of available methods for formal error propagation and simulation modelling. Special attention is given to the parameterization of different error models, i.e. how to make optimal use of existing information on error in the source data. When simulating error in the data layers spatial autocorrelation in the error field can be taken into account. This allows a more realistic assessment of uncertainty for spatial analysis which involves the application of neighbourhood or global operations.

---

## INTRODUCTION

The widespread availability and use of geographical information systems in recent years has greatly enhanced the opportunities for a more rationalized approach to spatial decision making. There is, however, a growing awareness that a successful use of GIS can only be achieved if it becomes possible to attach a quality label to the output of each GIS analysis. Therefore issues of spatial data quality, error modelling and error propagation have gained considerable interest from the GIS research community. The number of publications on the subject is rapidly increasing and several techniques for the modelling of uncertainty in spatial data sets as well as formal methods describing the mechanisms of error propagation have been developed.

Despite the impressive volume of research on the topic of error in spatial databases there is still a lack of proper tools for error handling in commercial GIS software. The aim of the work described in this paper is to bundle existing techniques for error modelling and error propagation in a software-independent tool. In designing the tool the following aims as stated by Heuvelink (1993) are pursued: 1. the error tool should not replace the GIS but only add functionality; 2. the error tool should be able to handle a variety of GIS operations; 3. the error tool must have a user-friendly interface to encourage its use; 4. the error tool should employ efficient routines to minimize the numerical load.

---

<sup>1,2</sup> Centre for Cartography and GIS, Department of Geography, Vrije Universiteit Brussel, Brussel, Belgium.  
E-mail: fforier@vnet3.vub.ac.be, fcanters@vnet3.vub.ac.be

The process of handling uncertainty in spatial data involves three steps. The first step consists in identifying for each data layer what kind of error information is available or can be extracted from other data sources. The type of error information will determine which assumptions have to be made when error is modelled (Lanter and Veregin 1992). For example, if thematic error is merely described by a global index of classification accuracy, e.g. the PCC index or kappa statistic, no thematic or spatial differentiation of error is possible. This means that one is forced to assume that error levels do not vary from class to class and are uniformly distributed over space, even although this will probably not be the case.

The second step in the error handling process involves the definition of a conceptual model of error. The choice which is made at this point will depend to a large extent on the kind of error information that is available (step 1). With respect to the type of error information used in the model a distinction can be made between global, thematic and spatial/thematic error models. The choice of an appropriate error model will also depend on the error propagation technique that will be applied (step 3). In this connection a distinction should be made between purely descriptive and stochastic error models.

Finally an appropriate error propagation function has to be selected to verify how error is propagated through the GIS operation. In the tool which is described in this paper two techniques for error propagation are considered, i.e. formal error propagation and simulation modelling. Formal error propagation uses explicit mathematical models which describe the mechanisms of error propagation for particular GIS operations (Veregin 1994). The formal approach has the advantage of easy applicability. Starting from a number of error indices which describe the quality of the data layer(s) involved in the operation a mathematical function is applied which directly calculates one or more error indices for the output layer. The problem with formal error propagation is that very little is actually known about error propagation mechanisms. Hence error propagation functions are only available for a small number of relatively simple GIS operations and prove to be rather restrictive with respect to the type of error models that can be used. If formal error propagation cannot be applied simulation modelling is the alternative. The main advantage of simulation modelling is its overall applicability regardless of the type of GIS operator which is involved. A disadvantage, however, is the high numerical load which makes that simulation modelling is less practical in connection with real-time error propagation.

## **IDENTIFICATION AND EXTRACTION OF ERROR INFORMATION**

One of the most important obstacles when handling uncertainty in GIS is the lack of knowledge about the error which is present in the source data. The type of error information which is available will define what kind of error modelling can be applied, whether error can be differentiated thematically and/or spatially, and as such will largely determine the quality of the whole error modelling process. Because the problem of error propagation in GIS has long been neglected information on the quality of the data is usually very sparse. Ideally, for each source layer in a GIS a map showing the spatial distribution of error should be available. Yet very often only global indicators of data quality are available providing no indication on local variations in the error field.

For categorical data information on error is usually described by a confusion matrix or classification error matrix. The matrix is constructed by cross-tabulation

of encoded and actual attribute values for a sample of locations. The element  $c_{ij}$  in the matrix represents the number of cells assigned to class  $i$  that actually belong to class  $j$ . A range of measures of accuracy such as the PCC (Proportion Correctly Classified) index and the kappa statistic (both global indices of classification accuracy), and the user's and producer's accuracy for each class can be deduced from the matrix (Congalton 1991). The confusion matrix reports error on the level of different thematic classes. Its major drawback is that it provides no information on within-class variations in the error field. The confusion matrix will therefore only support the application of thematically differentiated error models. Spatially differentiated models require extra information which is not contained in the matrix.

Categorical maps are usually not accompanied by a description of error in the form of a confusion matrix. When attempting to model error in GIS source layers that are derived from existing maps it is therefore often necessary to extract error information from a variety of other sources. Sometimes information on error can be found in reports which are delivered with map documents. Fisher (1991), for example, showed how information on soil map-unit inclusions from soil survey reports can be used to build a confusion matrix. Information on uncertainty can also be derived from field inventories (Aspinall and Pearson 1994) or from more detailed maps. Overlaying the original map with a more detailed spatial reference not only allows to derive a confusion matrix, but also makes it possible to obtain useful information on the spatial structure of uncertainty, for example, by calculating the mean size of inclusions of each class in the reference map within each class in the original source map. This kind of information is of utmost importance for the application of spatially differentiated error models, as will be shown later. Interesting information may also be provided by related spatial data. Fisher (1989) demonstrated how expert knowledge about the relation between soil and other physical data can be used to predict the spatial distribution of error in soil maps from the mapped distribution of these related variables.

In developing the tool which is referred to in this paper special effort has been made to provide the user with a set of functions which allow to derive error information for categorical maps from more detailed or related map documents. If no information on error is available or can be derived from other sources the user is offered the opportunity to introduce uncertainty in the data layers. This allows to carry out sensitivity analysis enabling the user to examine to what extent the output of a sequence of GIS operations is sensitive to specified levels of uncertainty in the source data.

## CHOOSING AN APPROPRIATE ERROR MODEL

Depending on the error information which is available three types of error modelling can be applied: global, thematic and spatial/thematic modelling. For global and thematic modelling the parameterization of the error models is based on well-known error indices that can be derived from the confusion matrix. Some error propagation techniques use the entire confusion matrix as a model of error (Veregin 1995). Spatial/thematic modelling requires additional information on the distribution of error within each thematic class.

In global error modelling thematic error is described by one single-valued index which is used as a model parameter. The best known error index that can be derived from the confusion matrix is the PCC (Proportion Correctly Classified) index which represents the probability that a location selected at random is correctly

classified. The PCC index is defined as the sum of the diagonal elements in the confusion matrix divided by the total number of samples. It has a value between 0 and 1, corresponding with maximum and no error respectively. Alternatively the kappa statistic may be used (Congalton 1991). The kappa index allows to account for change agreement and will be equal to zero if agreement is equal to what is expected in a random case. If agreement is less it will be negative, if agreement is perfect it will be equal to 1. The PCC index and the kappa statistic provide an overall measure of error. They do not describe how error may vary from class to class. Hence in global error modelling one is obliged to assume that error levels are the same for all classes and that confusion is equal among each pair of classes. If real error conditions largely deviate from these assumptions error propagation results based on global error modelling may be of little use.

Thematic error modelling accounts for variations in the level and type of error for different classes. Depending on the error propagation technique which will be used specific information on thematic error for different classes will be derived from the confusion matrix. For some GIS operations formal error propagation functions have been defined which propagate the entire confusion matrix of the source layer(s) through the operation to produce a confusion matrix for the output layer (Verigin 1995). This means that all error information which is present in the confusion matrix is actually used in the modelling. Hence the confusion matrix itself can be considered as the error model. Unfortunately at present formal error propagation functions are only available for a small set of simple GIS operations. If other operations are involved simulation modelling is the only alternative.

Simulation modelling requires the definition of a stochastic error model which allows to produce perturbed versions of the original data layer using available error information. Stochastic error models for categorical data randomly assign each cell in the data layer to one of the classes according to a priori defined class membership probabilities (Goodchild and Wang 1989, Fisher 1991). For all cells which have been assigned to a given class in the original source data membership probabilities for all classes can be derived from the confusion matrix by dividing the elements in the corresponding row of the matrix by the total number of sample locations that have originally been assigned to the class. Using the obtained class membership probabilities the most obvious way of introducing uncertainty into the source layer is to assign classes independently in each cell. As class membership probabilities for a cell always add to one each class can be matched to a sub-interval in the range 0-1 with its length proportional to the probability of that class. For each cell a random number in the range 0-1 is drawn and the cell is assigned to the class that corresponds with the interval including this random number. Applying a sequence of GIS operations to the perturbed data layer will yield an outcome that differs from the result obtained with the original source data. Differences in outcome which are obtained by repeating this process a large number of times will indicate how uncertainty is propagated through the analysis.

Whether formal error propagation techniques or simulation modelling is applied, the use of an error model which is solely based on class-specific information derived from the confusion matrix ignores the possible presence of autocorrelation in the error field. In fact a model which only permits thematic differentiation of error implicitly assumes that error distribution is purely random, in other words, that cells which are assigned to class *i* but actually belong to class *j* do not tend to cluster in space. In most cases this assumption will not be valid. For categorical data only one formal error propagation method is known to these authors which accounts for the presence of autocorrelation in the error field. The method allows to

estimate the accuracy of a buffer operation from a number of characteristics describing the source data and the error which is associated with it (Veregin 1994). Goodchild et al. (1992) presented a stochastic error model for categorical data which allows to control the level of spatial dependence by inducing correlation between outcomes in neighbouring cells. The mechanism of class assignment is again based on the cell's class membership probabilities. Yet instead of drawing a random number in the range 0-1 for each cell independently a spatially correlated random field is generated by means of a first order autoregressive model defined on a regular grid:

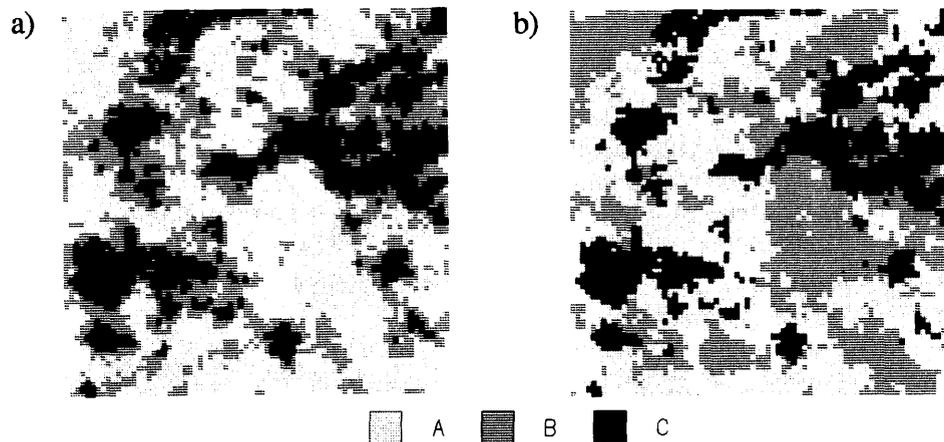
$$z_{i,j} = q (z_{i-1,j} + z_{i+1,j} + z_{i,j-1} + z_{i,j+1}) + e_{ij} \quad (1)$$

where  $z_{i,j}$  is the value of the random error field at position  $i,j$ ,  $q$  ( $< 0.25$ ) is a parameter that indicates the level of autocorrelation and where  $e_{i,j}$  is a spatially independent Gaussian noise field with mean 0 and standard deviation unity. Heuvelink (1992) presented a simple and robust iterative method for the generation of random fields of type (1) which avoids the computational overhead associated with the traditional inversion method (Haining et al. 1983). Since the marginal distribution of the correlated random field is of the Gaussian type with zero mean and a standard deviation which is function of the parameter  $q$  a simple transformation allows to obtain a uniform distribution in the range 0-1. This guarantees that class assignment for each cell is proportional to the a priori defined class membership probabilities.

Although it takes spatial correlation into account Goodchild's model has two important drawbacks. First of all, it is only applicable to two classes. If the model is used on more than two classes the inclusions which are simulated in the original map units will be spatially nested (figure 1a). Secondly there is no simple way of defining a proper value for the parameter  $q$ . Although the average size of patches in the simulated layer is clearly related to the value of  $q$  it is not possible to define a one-to-one correspondence between both as class membership probabilities disturb the relationship. Based on Goodchild's ideas a method is proposed which allows to simulate uncertainty in categorical maps with more than two classes. For each class in the source map the method takes into account the type and the mean size of inclusions as well as their proportion within the source map units. Instead of simulating inclusions of different classes simultaneously using one random field inclusions are generated sequentially (one class at a time), each time using a different random field. The advantage of sequential simulation is that adjacent inclusions will occur for each pair of classes, independent of the sequence of simulation (figure 1b).

The use of different random fields also allows to control the level of autocorrelation for each category and therefore provides a mechanism to adjust the size of inclusions. Unfortunately, as has been mentioned already, it is very difficult to define a clear relationship between the level of autocorrelation in the random field and the mean size of simulated inclusions. When sequential simulation is applied the relationship will not only be influenced by the class membership probabilities, yet also by the exact sequence of class simulation. As the level of autocorrelation which is required to generate inclusions of a given size cannot be defined a priori, an iterative method is used. Starting with a random field with a low level of correlation inclusions of one class are simulated. If the mean size of the inclusions is smaller than the proposed size the process is repeated with a random field which has a higher level of correlation, until the proposed size of inclusions is reached.

The major drawback of this procedure is its computational load since dealing with one class already involves the use of different correlated random fields. On-line generation of these fields would be impossible. That is why a large set of correlated random fields was generated for different levels of autocorrelation and permanently stored on disk. During the simulation process random fields with a required level of correlation are randomly picked from this set. At present the prototype tool for error modelling which has been implemented uses 15 sets of 50 correlated random fields, each set corresponding to a different level of autocorrelation. The fact that only 50 random fields are defined for each level has to do with limitations in storage capacity. The size of the fields was chosen large enough (1500 x 1500) to allow for their use in a large variety of applications.



**Figure 1.-** Simulation of classes B and C into class A with probabilities  $p_A = 0.5$  ,  $p_B = 0.3$  ,  $p_C = 0.2$  and level of autocorrelation  $q = 0.2450$  (a) simulation according Goodchild's model with 1 random field, (b) sequential simulation involving 2 random fields.

## METHODS FOR ERROR PROPAGATION

When source data are input to a GIS operation the uncertainty which is present in the data will be propagated through the operation and will introduce uncertainty in the outcome. Error propagation will continue when the output of one operation becomes the input to another. The network of input and output relations between spatial data layers can be schematized in a data flow diagram which is usually called a cartographic model (Tomlin 1990). Evaluating the accuracy of the output of a cartographic model is only feasible if the propagation of error can be modelled for each operation which is part of it. Although most GIS users may be aware that errors propagate through spatial analysis, in practice they will usually not pay any attention to it. An important reason for this is the lack of functionality for handling error propagation in commercial GIS.

Lanter and Veregin (1992) have proposed an automated system for error propagation that is based on the use of formal error propagation functions which transform the PCC index of the source layer(s) through the cartographic model of an application. The problem with formal error propagation methods, however, is that they are only available for a small number of GIS operations. Hence the use of tools which are strictly based on formal error propagation methods will be limited to

applications which only involve a few simple GIS operations. That is why some authors (Openshaw 1989) plead in favour of simulation modelling which puts no restrictions on the complexity of the application, thereby arguing that the computational load which is the major drawback of simulation modelling will become less a problem with the advent of increasingly powerful means of processing.

Although it may be questioned if the use of simulation modelling will allow enough runs to be performed in real-time to obtain a statistically valid quantitative assessment of uncertainty (Heuvelink 1993) it will give a clear indication of the level of uncertainty in the output which is to be expected when source data with known accuracy are subjected to a particular spatial analysis. While formal error propagation methods always produce a tabular report of error, i.e. a single error index or a confusion matrix, simulation modelling has the surplus advantage of providing interesting ways of visualizing uncertainty (see below). Hence it is the opinion of the present authors that most advantage can be derived from the complementary use of both error propagation techniques. In the following a short description is given of the techniques which have been implemented. Some reference is also made to their practical use.

### **Formal Error Propagation Methods**

Formal methods of error propagation can be defined as mathematical functions describing how errors in the source data are modified by a particular GIS operation. Each error propagation function is based on a set of assumptions about the nature of the errors present in the source data (a particular model of error) and is specific to one GIS operation. Assumptions about the nature of error will be based on the type of error information that is available for the source data, i.e. a global error index, class-specific indices of error or the entire confusion matrix. At present the prototype tool only includes functions which propagate the entire confusion matrix through the GIS operation. Functions are available for overlay (AND, OR, XOR) and for the reclass operation (Veregin 1995). If the entire confusion matrix is not known it is reconstructed from available error indices by making some simplifying assumptions about the distribution of error among different classes. So far no assumptions are made on thematic dependence, i.e. errors in different source layers are considered statistically independent.

For many GIS operations little is known about the mechanisms of error propagation and formal error models remain to be developed. Recently Veregin (1994) reported on a method to derive a formal mathematical expression describing how error is propagated through the buffer operation using simulation modelling. The method demonstrates how simulation modelling can add to a better understanding of error propagation mechanisms and to the development of formal expressions describing how error is propagated. It provides a good example of the complementary use of both error propagation techniques and of the benefits of integrating both in the same error modelling tool. Further research in this direction may lead to the development of formal error propagation functions for other more complex GIS operations.

Although the use of formal methods for error propagation in categorical data is still limited to a few GIS operations the technique is easy to apply. When performing an analysis where multiple source layers are subjected to a sequence of GIS operations it may be advantageous to combine simulation modelling with

formal error propagation in order to decrease the computational load. The combined use of both error propagation techniques has been a major concern in designing the tool which is described in this paper.

### **Monte Carlo Simulation Modelling**

Simulation modelling starts with the definition of a stochastic error model which allows to produce a set of perturbed versions of the original source data. As already explained the stochastic model will be based on a number of assumptions with regard to the type, the magnitude and the spatial distribution of error. The method for stochastic modelling which has been presented in the previous section and which allows to deal with autocorrelation in the error field has been implemented in the prototype version of the error handling tool. If no information on the spatial structure of error is available it is assumed that no correlation is present in the error field. If the mean size of inclusions in each class is known sequential simulation as described above can be applied.

Each perturbed version of the source data is subjected to the sequence of GIS operations involved in the analysis and will produce a different output. If the output of the analysis is a set of numerical values, e.g. the total area allocated to different types of land-use, the mean output over M realizations gives an estimate of the most likely outcome of the analysis taking the uncertainty in the source data into account. The standard deviation over M realizations provides a measure of the uncertainty in the output. If the outcome of the analysis is an entire coverage, e.g. a visibility map, simulation results allow to generate a variety of products describing output uncertainty (see below).

The implementation of the stochastic modelling procedure is based on scripting. When the analysis is executed for the first time the cartographic model is stored in a script file. Monte Carlo analysis is automatically performed by running the script a fixed number of times. This strategy, of course, assumes that the GIS allows scripting and that the script language includes control structures for iteration.

### **VISUALIZATION OF UNCERTAINTY**

An important issue in the development of a tool for error handling is how to visualize the quality of the source data and the reliability of the output of spatial analysis in an easy and straightforward manner. To give the user a visual impression of the error present in the source data it may be interesting to visualize the result of one stochastic simulation of error. When dealing with classified satellite data class membership probabilities produced by statistical classification techniques can be used to visualize uncertainty which is due to the classification process. Based on these probabilities not only the first, but also the second, third, ... most likely class can be shown. To evaluate the reliability of pixel assignment a probability image can be constructed which shows the highest class membership probability for each pixel. Confusion in pixel assignment can be measured by calculating the relative probability entropy (Maselli et al. 1994). This index indicates for each pixel if probability tends to concentrate in one class or is spread over several classes.

When error is propagated using the Monte Carlo approach and the outcome of the analysis is an entire coverage the same techniques as described for the visualization

of uncertainty in classified satellite data can be applied since for each pixel a vector indicating the frequency of occurrence of each class can be constructed from the M coverages. If error is propagated using formal error propagation methods uncertainty cannot be visualized since output is in tabular form.

## **IMPLEMENTATION**

The above described techniques for error modelling and error propagation have been implemented in the C programming language in a UNIX workstation environment. The prototype version of the tool is linked to GRASS 4.1. The choice of GRASS has been based on the fact that it is public-domain software, written in C and running under all common UNIX operating systems and that it provides a powerful language for scripting. The most important advantage of GRASS, however, is its open structure, meaning that developers can directly add new functionality to the system. The approach that has been followed during the implementation of the error handling tool was to consider the GIS as an external module which is used to carry out spatial analysis and to display maps. Hence the tool operates independently from the GIS and should be easy-to-link with other GIS software.

## **CONCLUSION**

This paper reviews existing methods for error modelling and error propagation in raster-formatted categorical data. The aim of the work reported here was to bundle these methods in a user-friendly and independent tool for error handling which can easily be linked with existing GIS software. A prototype version of the tool has been implemented using GRASS 4.1 for raster-based spatial analysis and display.

The tool combines the use of formal error propagation methods with simulation modelling. Both techniques should not be considered as two opposites but rather as complementary techniques. When spatial analysis involves the use of GIS operations for which no formal error propagation function exists simulation modelling can be applied. At present formal methods have been implemented for overlay analysis (AND, OR, XOR) and for the reclass operation. They allow to propagate the entire confusion matrix through the operation. Simulation modelling is based on a method for sequential simulation which allows to deal with spatial autocorrelation in the error field. In the implementation of the method specific measures were taken to reduce computational load.

## **ACKNOWLEDGMENTS**

The work reported in this paper is supported by the Flemish Institute for Scientific and Technological Research for the Industry (IWT) and by the Belgian Scientific Research Programme on Remote Sensing by Satellite - phase three (Federal Office for Scientific, Technical and Cultural Affairs), contract Telsat/III/03/004. The scientific responsibility is assumed by its authors.

## REFERENCES

- Aspinall, R.J. and D.M. Pearson, 1994. A method for describing data quality for categorical maps in GIS, Proc. of the Fifth European Conference and Exhibition on Geographical Information Systems, EGIS/MARI'94, Paris, France, 444-453.
- Congalton, R.G. 1991. A review of assessing the accuracy of classifications of remotely sensed data, *Remote Sensing of Environment*, 37, 35-46.
- Fisher, P.F. 1989. Knowledge-based approaches to determining and correcting areas of unreliability in geographic databases. Accuracy of spatial databases, eds: Goodchild, M. and S. Gopal, 45-54.
- Fisher, P.F. 1991. Modelling soil map-unit inclusions by Monte Carlo simulation, *Int. J. of Geographical Information Systems*, 5, 2, 193-208.
- Goodchild, M.F. and M.H. Wang, 1989. Modelling errors for remotely sensed data input to GIS, Proc. on the Ninth International Symposium on Computer-Assisted Cartography, AUTO-CARTO 9, Baltimore, Maryland, 530-537.
- Goodchild, M.F., S. Guoqing and Y. Shiren, 1992. Development and test of an error model for categorical data, *Int. J. Geographical Information Systems*, 6, 2, 87-104.
- Haining, R.P., D.A. Griffith and R.J. Bennett, 1983. Simulating two-dimensional autocorrelated surfaces. *Geographical Analysis* 15, 247-255.
- Heuvelink, G.B.M. 1992. An iterative method for multidimensional simulation with nearest neighbour models, in Dowd, P.A. and J.J. Royer (eds.), 2nd CODATA Conference on Geomathematics and Geostatistics, Nancy, Sciences de la Terre, Series Informatique et Geologique, 31, 51-57.
- Heuvelink, G.B.M. 1993. Error propagation in quantitative spatial modelling, applications in geographical information systems, Netherlands Geographical Studies, KNAG, Faculteit Ruimtelijke Wetenschappen Universiteit Utrecht, 151 p.
- Lanter, D.P. and H. Veregin, 1992. A research paradigm for propagating error in layer-based GIS, *Photogrammetric Engineering & Remote Sensing*, 58, 6, 825-833.
- Maselli, F., C. Conese and L. Petkov, 1994. Use of probability entropy for the estimation and graphical representation of the accuracy of maximum likelihood classifications, *ISPRS J. of Photogrammetry & Remote Sensing*, 49, 2, 13-20.
- Openshaw, S. 1989. Learning to live with errors in spatial databases. Accuracy of spatial databases, eds: Goodchild, M. and S. Gopal, 263-276.
- Tomlin, C.D. 1990. *Geographic Information Systems and Cartographic Modelling*. Englewood Cliffs: Prentice Hall.
- Veregin, H. 1994. Integration of simulation modelling and error propagation for the buffer operation in GIS, *Photogrammetric Engineering & Remote Sensing*, 60, 4, 427-435.
- Veregin, H. 1995. Developing and testing of an error propagation model for GIS overlay operations, *Int. J. of Geographical Information Systems*, 9, 6, 595-619.

## BIOGRAPHICAL SKETCH

Both authors are research assistants in the Department of Geography of the Vrije Universiteit Brussel, Belgium. Their main research interests lie in spatial database accuracy, error modelling and error propagation in GIS.

# Discrete Polygons or a Continuous Surface: Which is the Appropriate Way to Model Forests Cartographically?

Kim Lowell<sup>1</sup>

---

**Abstract.**--Maps traditionally represent forests as a group of discrete polygons -- i.e., forest stands. However, as GIS technology has become more common, there is an increasing use of interpolation techniques to represent forest parameters as a continuous surface. One must question if both can be correct. This paper presents the results of work conducted to assess the validity of a polygon vs. continuous surface representation of forest volume. Specifically, for two different sample densities, spatial autocorrelation was measured among sample plots. Following this the improvement in volume estimation provided by spatial interpolation for individual plots was assessed relative to the mean and standard deviation for all plots. It is concluded that in virtually no situation -- between forest types, different stands of the same forest type, and within a single stand -- is it appropriate to model forest volume as a continuous surface. Conversely, the evidence suggests that the forest is a set of polygons with inexact boundaries.

---

## BACKGROUND AND CONTEXT

### Spatial Interpolation

Spatial interpolation is a technique which has existed since long before digital geographic information systems (GIS) were created. It is used to obtain estimates over an entire surface for a given variable which has only been sampled at isolated points. Common examples are estimating rainfall and/or temperature over an entire region for which weather data were collected at isolated meteorological stations. But even this seemingly simple and straightforward example masks a number of potential problems with its use in such a situation. All are related to two underlying implicit assumptions which must be satisfied if the resulting interpolation is to be a valid representation of the phenomenon of interest.

First, the phenomenon being interpolated must truly be continuous. Suppose that one wishes to model rainfall for a region only during a summer period of hot, humid weather. In such conditions, it is likely that the majority of precipitation will be convective producing considerable amounts of rain in spot locations rather than a smooth gradient of rainfall over a large region. If this is the case, then rainfall for the summer period may not be an interpolable phenomenon due to the seemingly random nature of convective precipitation.

<sup>1</sup>*Professeur titulaire, Centre de recherche en géomatique, Université Laval, Pavillon Casault, Ste-Foy, Québec G1K 7P4. Kim.Lowell@for.ulaval.ca. (418) 656-7998. FAX: (418) 656-7411.*

Second, while the initial necessary condition for interpolability is a continuous phenomenon, the second important factor is that the phenomenon being studied must be sampled at an appropriate density for its continuous nature to manifest itself. Suppose that in another region, all precipitation is orthographic -- an interpolable phenomenon which is strongly related to the topography of a region. Further suppose that in the region storm fronts arrive from the west and, as they move to the east, encounter a series of mountain summits of increasing elevation evenly spaced 10 km apart. It is reasonable to believe that the western slopes of these mountains will receive increased rainfall as one moves towards a summit. Moreover the amount of rainfall will fall off sharply as a summit is crested given that a storm rises in contact with a mountain slope, cools, releases all possible precipitation, and moves towards the next higher mountain where the process repeats itself. In this hypothetical case, if weather stations are spaced, say, 40 km apart, the resulting data for rainfall may not lend themselves to being interpolated even though the phenomenon being studied is known to be interpolable. A better approach might be a regression model in which the amount of rainfall is estimated as a function of elevation and west- or east-facing slope.

Thus the ability to conduct reliable spatial interpolation depends both on the nature of the phenomenon being studied and also the way in which it has been sampled.

### **Forest Parameters**

Traditionally forest maps are composed of non-overlapping polygons having non-overlapping appellations within the classification system used. Yet recently -- particularly with the advent of GIS and associated technologies -- interpolation has begun to be used to produce continuous surfaces from forest sample points for various forest parameters.

And, indeed, there may be reason to believe that at some scale forest parameters are interpolable phenomenon. Given a single forest stand, for example, it may be that "a lot" of its volume is concentrated in a relatively small area with gradations towards adjoining forest stands. Some researchers (Joy and Klinkenberg 1994) have noted the tendency for forest polygon cores to be the most precise -- i.e., repeatable by several interpreters -- part of a polygon suggesting that forest parameters may show a gradient from polygon centre to polygon edges. Such information has led to at least one representation of spatial certainty for various forest types based on the idea that the polygon centre is the part of the polygon which is most likely to truly be the forest type labelled -- an interpolation of certainty (Lowell 1994). And between forest stands of the same type or even different forest types it may be possible that forest parameters are interpolable. A pure spruce stand in northern Quebec may contain less volume than the same stand type in southern Quebec. Similarly, a high volume stand in one location may suggest that surrounding stands will also have high volumes due to favourable conditions related to underlying factors such as soil type and water availability.

The converse may also be true, however. Consider that the spatial structure of within-stand variance is not known and would seem to depend strongly on the forest sampling unit employed -- a smaller sample unit giving a larger variance than a larger unit. Moreover, in certain stands the spatial variability of tree locations and the characteristics of the sample units employed may render the spatial structure of the variance seemingly random -- i.e., non-interpolable. Similarly, if a forest has been subjected to periodic disturbances such as fires or severe wind, the characteristics of one stand may provide absolutely no indication of the characteristics of nearby

stands. Moreover, even different forest stands within the same forest type may show no gradation of productivity. Over a long range -- i.e., northern to southern Quebec -- it may not be possible to have the same forest type appear over such a range due to biological carrying capacities of various sites. And even over a smaller region the productivity of the stands of a given forest type may be more related to local ecophysiological conditions than to a mere gradation of productivity.

## Objectives

The purpose of this paper is to examine the question of whether or not forest volume is an interpolable phenomenon in the boreal forest of Quebec. This question will be addressed at a number of levels as it concerns volume among forest stands of different forest types, different forest stands in the same forest type, and within individual forest stands. In addition to assessing whether or not volume is interpolable at any of these scales, the implications of conducting interpolation when it is not warranted for these conditions will also be evaluated and presented.

## STUDY AREA, DATA, AND METHODS

Two real world forestry data sets were employed in this study. Both contain the per hectare wood volumes estimated for a collection of 0.04 ha sample plots located in Montmorency Forest -- the research forest of Laval University located 80 km north of Quebec City. The dominant vegetation on Montmorency Forest is the boreal forest dominated by balsam fir (*Abies balsamea* (L.) Mill.) and to a lesser extent spruce (*Picea* spp.) and white birch (*Betula papyrifera* Marsh.) (Bélanger *et al.* 1988). The first data set is composed of 92 sample plots distributed in a semi-random stratified sample throughout the 1500 ha study area. These plots had been established as part of normal management activities to assess the harvestable volume on the forest. They were therefore located in mature stands or those approaching a harvestable volume. The density of these plots (1 per 16 ha or 1 plot every 400 m on average) is approximately 20 times as dense as forest inventories normally taken in Quebec (Beaulieu and Lowell 1994). The second data set contains 81 plots located within a 16 ha square in Montmorency Forest. These plots were placed on a 9-by-9 grid with plots being located 50 m apart. The 16 ha grid was located intentionally such that a number of forest types covering a range of forest conditions were included in the sample: clearcuts, partial and total windthrow, and mature and developing forests.

Two types of analyses were conducted on each data set in its entirety and a series of subsets. These subsets will be described subsequently.

First, to assess the interpolability of volume, the spatial autocorrelation was measured among the plots of a given data (sub)set (see Table 1). This was assessed by using both Moran's  $I$  and Geary's  $c$  (Griffith 1985) and the assumption that sample plots were adjacent if, in a Voronoi diagram constructed around the plots, their Thiessen polygon's touched. If a phenomenon is interpolable, it is expected that high values will have a tendency to cluster in the same area as will low values. This is the same as saying that positive spatial autocorrelation will be present.  $T$  values for  $I$  and  $c$  are presented to indicate significant differences from zero. If forest volume is interpolable, it is expected that these values will be positive and significant.

Second, a given data (sub)set was divided into a calibration and validation subset. The validation subset contained a random sample of interior plots that was

approximately one-third of all the interior plots. The calibration data set contained the remaining plots. Spatial interpolation using Voronoi diagram area stealing (Gold 1989) was conducted on the calibration data set and the interpolated volumes for each of the validation data plot locations extracted. The mean difference of the true and interpolated values was calculated and tested for a significant difference from zero. The two values were also regressed against each other; perfect agreement will provide an  $R^2$  of 1.0, a slope of 1.0 and an intercept of 0.0, a root-mean-square-error (RMSE) of 0.0.

**Table 1. Data sets and subsets treated, motivation for subset, and number of points in the validation and calibration subsets of each.**

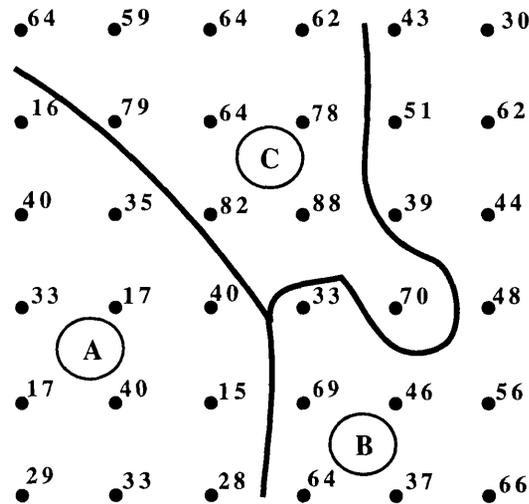
<b>Data (Sub)set</b>	<b>Motivation</b>	<b>Points: Total Cal Val</b>		
<i>1500 ha (Between stand interpolability)</i>				
All plots	Global interpolability	92	77	15
Softwood Plots	Interpolability within species	81	69	12
Fir plots $\geq 60\%$ density	Interpolability of fully stocked stands	61	52	9
<i>16 ha (Within stand interpolability)</i>				
All plots	Global interpolability	81	67	14
$\geq 250$ stems per ha	Across species types w/o clearcuts and blowdowns	65	54	11
Softwood and $\geq 250$ stems per ha	Within a single (dominant) species type	55	46	9
<i>Benchmarks</i>				
Topographic	Known interpolable phenomenon	1037	998	39
Random	Non-interpolable phenomenon	100	81	19
Synthetic data	Explanation of results	36	29	7

The subsets of the two real world forestry data sets examined were designed to test the interpolability of forest volume at a variety of levels under the conditions studied (Table 1). For the 92 plots over 1500 ha the subsets were a)all data together, b)only plots labelled as softwood on the forest map of the area, and c)only plots labelled predominantly fir with a density greater than 60%. For the 81 plots on a 16 ha grid the subsets were a)all data together, b)only plots having 250 stems per ha and higher, and c)only softwood plots having 250 stems per ha or higher.

For a point of comparison, three additional data sets were also treated and analysed in exactly the same manner as the other data (sub)sets. First, a digital terrain model for a 1 km (100 ha) square developed for another study was employed (Lowell 1994). Second, 100 locations were randomly generated over a 500 m square area (25 ha) and a value between 1 and 100 (with replacement) randomly assigned to each of the locations. These two data sets provide benchmarks for what one should find for a phenomenon known to be interpolable (elevation) and one which is not at all expected to be interpolable (random locations and attribute assignments). The third synthetic data set was a grid of 36 points spaced 50 m apart (Figure 1). The use of this data set will become clear after the results of the other data sets have been presented and discussed. For the moment, suffice it to say that this data set is designed to simulate three forest types which were sampled systematically and densely. The value assigned at each location was random within the range assigned *a priori* to each type: A) 10 to 40, B) 30 to 70, and C)60-90. The overlap between classes is intentional as it reflects real forest conditions wherein it is possible that the volume at any given location for a particular type is less than that found in a type

known to contain less volume. The mean, of course, of the two types will reflect their relative densities.

**Figure 1. Synthetic forest stand. Ranges for types: A)10 to 40, B)30 to 70, C)60 to 90.**



## RESULTS AND DISCUSSION

The results for the two benchmark data sets provide a useful manner in which to assess the results for the real world forest data sets (Table 2). In general, both the topographic and the random data sets perform as expected.

The topographic data are strongly positively spatially autocorrelated suggesting that interpolation for this phenomenon at the scale sampled is warranted. This is further reinforced by noting the strong relationship that exists between the true and interpolated data for the validation data points ( $R^2=0.97$ ), and the mean difference between the two sets of values is fairly small. Moreover, the slope and intercept of the regression line are close to the expected values of 1.0 and 0.0, respectively. Another measure of the usefulness of interpolation can be seen in the value of the RMSE (2.5) compared to that of the standard deviation (14.1) for all data. In the absence of the interpolated values, the best estimate for a given point's elevation will be 60.1 (the overall mean) with the error varying  $\pm 28$  with 95% confidence (a 95% confidence limit using the standard deviation under an assumption of a normal distribution). The interpolation will provide an estimate at a given location that varies only  $\pm 5$  with 95% confidence.

By contrast, the random data set indicates no spatial autocorrelation structure (non-significant  $I$  and  $c$ ) suggesting that spatial interpolation is not warranted. And indeed the  $R^2$  indicates that no correlation exists among the interpolated and true values for the validation data points. Furthermore the slope of the line is significantly different from the expected value of 1.0 though not significantly different from a slope of zero (which indicates no relation). In the absence of a useful interpolation, the value estimated at a given point will be the mean ( $52.6 \pm 56.1$ ) with 95% confidence while the interpolation gives estimates within 61.9 units with 95% confidence. Clearly, the random values should not have been interpolated as indicated by the spatial autocorrelation statistics and the subsequent regression analysis. In fact, the

consequences of having conducted an interpolation when it was not warranted has been in increase in the error on the resulting surface. In a practical sense, this means that the resulting surface has virtually no accuracy for a given location.

**Table 2. Results of analysis for spatial autocorrelation and interpolation.**

Data set	Mean <sup>1</sup> Spat. Auto. <sup>2</sup>			Regression <sup>3</sup>				All Data <sup>4</sup>	
	Diff.	I	c	R2	Slope	Intercept	RMSE	Mean	Std Dev.
<i>1500 ha (Between stand interpolability)</i>									
All plots	-18.6	0.98	0.62	0.16	0.95	-12.5	65.5	118.6	62.8
Softwood Plots	36.5	0.97	0.93	0.00	-0.38*	139.1	70.1	116.8	63.8
Fir plots ≥ 60% density	-30.8	0.53	0.41	0.00	0.08**	90.0	62.1	121.3	62.2
<i>16 ha (Within stand interpolability)</i>									
All plots	2.5	4.02**	3.56**	0.00	0.14	66.9	41.7	73.7	43.2
≥ 250 stems/ha	2.1	3.25**	1.88	0.00	0.38	54.0*	22.3	89.5	31.6
Softwood and ≥ 250 stems/ha	-11.9	1.77	0.81	0.00	-0.24	97.3**	28.2	92.6	31.7
<i>Benchmarks</i>									
Topographic	-0.1	52.26**	47.36**	0.97	1.04	2.4	2.5	60.1	14.1
Random	-5.9	-1.01	-0.38	0.00	0.18*	39.6	31.6	52.6	28.6
Synthetic Data	-1.3	3.16**	2.02*	0.07	1.22	-12.7	28.2	48.4	19.9

<sup>1</sup>Validation data only. "\*"/\*\*" indicates significant difference from zero at 95%/99%.

<sup>2</sup>t values for Moran's I and Geary's c. for all data. "\*"/\*\*" indicates significant difference from zero at 95%/99%.

<sup>3</sup>Validation plots only. "\*"/\*\*" indicates significant difference from expected values at 95%/99%.

<sup>4</sup>All data for purposes of comparison.

It is also apparent that none of the situations explored using the 92 sample plots are interpolable. Perhaps the quickest way to see this is to note that none of the interpolations provided RMSEs that were noticeably different from the standard deviations of the data. However, this is also abundantly clear in the low R<sup>2</sup>s, the coefficients of the regression lines, and the lack of significance of either spatial autocorrelation coefficient. The importance of these findings is considerable.

This demonstrates that, in the conditions studied, between-stand interpolation is not warranted. This is perhaps not surprising in the case of global interpolability (All plots) and interpolability within species (Softwood plots) considering that Montmorency Forest is a managed forest on which normal harvesting activities have been ongoing. Thus the volume of the stands represented in these two cases may reflect more on past management practices than the biological productivity of these areas -- even though plots were located in relatively mature forests. Note that these conclusions give little idea about what would be found in forests which have not been subjected to wide-spread natural or human disturbances. However, it is argued here that such forests are not very common meaning that these results should be widely applicable.

The findings for the 92 plots also indicate that one cannot interpolate even for fully stocked stands of similar species combination which have not been subjected to recent disturbance. These stands (Fir plots ≥ 60% density) should have had time to manifest their biological potential and possibly show a gradient of productivity across the roughly 7 km by 3 km study area. It might be suggested that the reason

that this was not the case was that the productivity of each stand is related to the underlying ecophysiology of a stand which was not considered here. However, in another study which used the same 92 sample plots (Coulombe and Lowell 1995) the relation of the productivity of each site to its underlying ecophysiology as estimated by map-based data (soils, slope, etc.) was examined; no significant relationships were found. Thus it would appear that over a large area, even with a sampling density much higher than is usually employed, interpolation of forest volume is not appropriate even for stands which have remained undisturbed for a long period.

At first glance, the results for the various subsets of the 81 sample plots might indicate that within-stand interpolation is possible -- both spatial autocorrelation coefficients are positive and relatively large for All plots and plots having  $\geq 250$  stems per ha. However, the associated  $R^2$ 's are zero (0.0) and the RMSEs are not largely different from the standard deviations. A visual examination of the data makes it very clear that the reason the spatial autocorrelation statistics are strong in the case of All Plots is the presence of four adjacent plots being located in a clearcut, and another five plots being located in a blowdown area. Simply stated, low volumes cluster in these areas -- i.e., are positively spatially autocorrelated. When these are removed, spatial autocorrelation decreases, though it remains positive and strong. While this suggests that interpolation would be appropriate, the regression statistics suggest otherwise.

This can be explained by considering the synthetic surface in Figure 1 and its associated statistics in Table 1. This surface was constructed so that large values would have a tendency to cluster (which is the case as measured by the spatial autocorrelation coefficients for this surface). But if one removes a given observation and attempts to estimate its value from surrounding observations, the precision of estimates is fairly low. That is, like-values show tendencies to group, but within each group there is no consistency of estimates. This suggests that there are identifiable areas -- i.e., polygons -- of high and low values, but within those areas, the variance is spatially random.

That this is the case in real forests is further reinforced by considering the Softwood and  $\geq 250$  stems per ha plots. In this data subset, spatial autocorrelation is weakened to the point where it is no longer statistically significant, and the regression statistics show that interpolation has not provided better estimates than would the mean and the standard deviation alone. Given that the case of Softwood and  $\geq 250$  stems per ha can be considered a single type, if one cannot observe a spatial tendency of volumes within a stand then one should not conduct within-stand interpolation.

To understand all of the findings of this study -- i.e., a complete lack of interpolability -- one must consider the nature of forest sampling. With conventional forest inventory techniques, the volume found at any given location is somewhat a function of the micro-location of the plot. That is, each forest type has an associated range of volume that one will find at any given location. When one samples a given location, one cannot know if one is at the upper, lower, or middle portion of that range. Thus a large part of the reason that between- and within-stand interpolability is inappropriate is simply the natural random spatial variability of a forest type and the way in which it is sampled. One could minimise such effects by using a larger sample unit so that one is consistently sampling the mean of a type (or the range for the type would be reduced). But doing so risks to require an excessive amount of field time and also to exceed the boundaries of the forest stand being sampled.

## CONCLUSIONS

It appears that it is not appropriate to model the forest as a continuous surface. Instead, the results from this study suggest that the forest is a set of polygons. Note, however, that this does not mean that one has polygons with crisp boundaries as is usually represented on forest maps. While the synthetic surface employed herein (Figure 1) was based on definite, widthless boundaries, a number of different interpreters presented with only the 36 points in Figure 1 would probably not develop the same map. And even if such interpreters were provided with the knowledge that there were only "truly" three types present, it is doubtful that the lines would be placed in the same locations by different interpreters. Thus the best interpretation of how to represent a forest might be as a set of polygons whose boundaries show a range of possible locations.

## ACKNOWLEDGEMENTS

The financial support for this work provided by the Association of Quebec Forest Industries and the Natural Sciences and Engineering Research Council of Canada is gratefully acknowledged.

## BIBLIOGRAPHY

- Beaulieu, P., Lowell, K., 1994. Spatial autocorrelation among forest stands identified from the interpretation of aerial photographs. *Landscape and Urban Planning* **29**:161-169.
- Bélanger, L., Bertrand, L., Bouliane, P., Lussier, L.J., 1988. *Plan d'aménagement de la forêt Montmorency*. Université Laval, Faculté de foresterie et de géomatique, 215 pp.
- Coulombe, S., Lowell, K., 1995. Ground-truth verification of relations between forest basal area and certain ecophysiological factors using a geographic information system. *Landscape and Urban Planning* **32**:127-136.
- Gold, C.M., 1989. Surface interpolation, spatial adjacency and G.I.S. Chapter 3 in *Three dimensional applications in geographic information systems* (J. Raper, Editor), Taylor and Francis, London, pp. 21-35.
- Griffith, D.A., 1985. *Spatial autocorrelation: a primer*. Monograph of the American Association of Geographers, Washington, 82 pp.
- Joy, M., Klinkenberg, M., 1994. Proceedings: *GIS '94*, February, Vancouver, British Columbia, pp. 359-365.
- Lowell, K.E., 1994. Probabilistic temporal GIS modelling involving more than two map classes. *International Journal of Geographical Information Systems* **8**:73-93.
- Lowell, K.E., 1994. fuzzy cartographic representation for forestry based on Voronoi diagram area stealing. *Canadian Journal of Forest Research* **24**:1970-1980.
- Odland, J., 1988. *Spatial Autocorrelation*. Scientific Geography Series, Volume 9, Sage Publications, Beverly Hills, California. 87 pp.

## BIOGRAPHICAL SKETCH

Kim Lowell is a Full Professor of Forestry Geomatics at Laval University in Quebec City, Canada within the Industrial Research Chair in Geomatics Applied to Forestry. He received a Ph.D. in Forest Biometrics from Canterbury University, New Zealand in 1985.

# Choosing between Abrupt and Gradual Spatial Variation?

Gerard B.M. Heuvelink<sup>1</sup> and Johan A. Huisman<sup>2</sup>

---

**Abstract.**-Two basic models of spatial variation are widely used in present-day soil survey practice. The discrete model of spatial variation (DMSV) forms the basis of the traditional soil map, in which homogeneous soil mapping units are separated by abrupt boundaries. The continuous model of spatial variation (CMSV) originates from geostatistics, where kriging is used to map gradual changes in soil properties. Neither of the two models is capable of handling situations in which abrupt and gradual spatial variation are both present in the same area. Therefore, recently a straightforward combination of the two models was introduced, known as the mixed model of spatial variation (MMSV). The MMSV contains the two basic models, suggesting that it may perform well on the whole range of spatial variation. In this paper we investigate the anticipated flexibility of the MMSV using a simulation study. As expected, the MMSV is superior to the DMSV and the CMSV when both types of spatial variation are present. But the MMSV is also as suitable as the DMSV in case of a 'discrete' reality, and as suitable as the CMSV in case of a 'continuous' reality. From this we conclude that the MMSV should be recommended for situations where an a priori choice between abrupt and gradual spatial variation cannot easily be made.

---

## INTRODUCTION

Traditional soil survey is based upon the presumption that soil behaves uniformly within soil mapping units and changes fairly abrupt at the boundaries between them (Voltz and Webster 1990, Webster and Oliver 1990, Burrough 1993). However, the practical validity of this conventional representation of soil variability has repeatedly been questioned (e.g. Webster and Cuanalo 1975, Nortcliff 1978, Campbell et al. 1989, Nettleton et al. 1991). Major drawbacks of the conventional model of soil spatial variation are that it cannot represent gradual boundaries and that it ignores spatial autocorrelation within mapping units.

As an alternative to the conventional model, some fifteen years ago the use of geostatistical techniques for modelling spatial variation was introduced to soil science. Burgess and Webster (1980) were among the first to apply kriging, an

---

<sup>1</sup>Geostatistician, University of Amsterdam, The Netherlands

<sup>2</sup>Student, University of Amsterdam, The Netherlands

exponent of this theory, to soil survey. Since then the geostatistical approach to modelling soil spatial variation has flourished and kriging is now routinely being adopted for the mapping of soil properties (Webster 1994).

Recently, however, it is more and more being realized that the outright abandoning of the conventional approach to soil spatial variation is perhaps too drastic. Kriging definitely has disadvantages as well, such as its inadequacy to deal with sharp boundaries. In order to bridge the gap between the conventional and geostatistical representation of soil spatial variation, several models have been developed that can handle discrete (abrupt) as well as continuous (gradual) spatial variation in the same area (Stein et al. 1988, Voltz and Webster 1990, Heuvelink and Bierkens 1992, Rogowski and Wolf 1994, Goovaerts and Journel 1995, Heuvelink 1996).

In this paper we examine one such a model, known as the mixed model of spatial variation (MMSV). In Heuvelink (1996) it was anticipated that the MMSV should work well on the whole range of spatial variation, from purely discrete to purely continuous. We analyze the anticipated flexibility of the MMSV using nine simulated 'realities'. But before describing the exact procedure of the simulation exercise, we will first briefly review the three models of spatial variation used in this study.

### THREE MODELS OF SPATIAL VARIATION

The *Discrete Model of Spatial Variation* (DMSV) first divides the geographical domain  $D$  into  $K$  separate units  $D_k$ . It then makes the following assumptions on the behaviour of a spatially distributed attribute  $Z(\cdot)$ :

- 1  $Z(x) = \mu_k + \epsilon(x)$  for all  $x \in D_k$
- 2  $\epsilon(\cdot)$  has zero mean and is spatially uncorrelated
- 3  $\text{Var}(\epsilon(x)) = C_0$  for all  $x \in D$

Thus the DMSV assumes that  $Z(x)$  is the sum of a unit-dependent mean  $\mu_k$  and a residual noise  $\epsilon(x)$ . The DMSV will usually be adopted when the units  $D_k$  are available in the form of a polygon map, such as a soil map, a landuse map or a geological map, and when the within-unit variability is expected to be small in comparison with the between-unit variability. In other words, the DMSV represents the conventional model of soil spatial variation and is appropriate when major jumps in the attribute  $Z(\cdot)$  take place at the boundaries of the mapping units.

In its simplest form, the *Continuous Model of Spatial Variation* (CMSV) makes the following assumptions:

- 1  $E[Z(x)] = \mu$  for all  $x \in D$
- 2  $\text{Cov}(Z(x), Z(x+h)) = C_z(|h|)$  for all  $x, x+h \in D$

Thus the CMSV assumes that  $Z(\cdot)$  is second-order stationary, meaning that it has a constant mean and that its spatial autocovariance is a function only of the distance between the locations. The CMSV embodies the geostatistical model of spatial variation. In geostatistics it is customary to use the variogram  $\gamma_Z(\cdot)$  to characterize the spatial autocorrelation of  $Z(\cdot)$ . It is related to the autocovariance by the identity  $\gamma_Z(|h|) = C_Z(0) - C_Z(|h|)$ .

The assumptions underlying the *Mixed Model of Spatial Variation* (MMSV) are a combination of those underlying the DMSV and CMSV:

- 1  $Z(x) = \mu_k + \epsilon(x)$  for all  $x \in D_k$
- 2  $\epsilon(\cdot)$  has zero mean
- 3  $\text{Cov}(\epsilon(x), \epsilon(x+h)) = C_\epsilon(|h|)$  for all  $x, x+h \in D$

Second-order stationarity is thus imposed on  $\epsilon(\cdot)$  instead of on  $Z(\cdot)$ . The MMSV is more general than the DMSV and CMSV, and in fact it contains both models. However, the DMSV and CMSV are included here as separate models because they are very often used in practice.

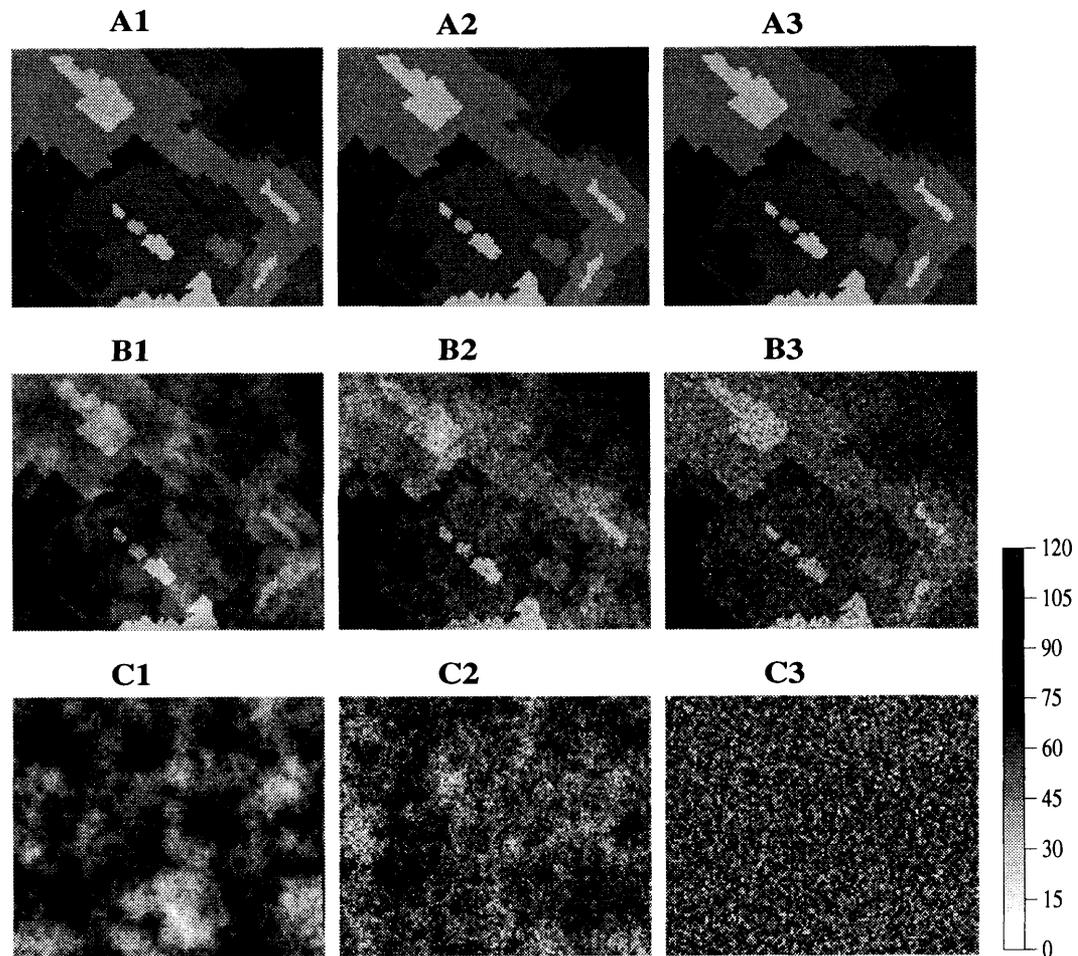
### APPLICATION TO NINE SIMULATED REALITIES

In order to study the suitability of the three MSV's for mapping under various circumstances, nine different 'realities' were created. This was done by adding maps generated using unconditional Gaussian simulation (Deutsch and Journel 1992) to an artificially constructed 'soil map'. The nine realities are given in figure 1. The A-maps in figure 1 (top row) are strongly dominated by the discrete soil map, the B-maps (middle row) to a much lesser extent, and the C-maps (bottom row) bear no influence from the discrete soil map. The degree of spatial autocorrelation in the added residual decreases from the 1-maps (left column) to the 3-maps (right column).

#### Mapping the soil property from 200 observations

From each of the nine simulated maps data sets were created by collecting observations at 200 randomly selected locations. From these observations the soil property was mapped using the three MSV's. This resulted in 27 maps of predictions and 27 maps of prediction error standard deviations.

Mapping using the DMSV is done simply by calculating the mean of the observations for all mapping units separately, and using the unit mean as a prediction for all points lying in the same unit. Mapping with the CMSV is done by ordinary kriging, in which a variogram is used computed from the 200 observations. Mapping is somewhat more complicated in case of the MMSV. First a variogram is computed from the 200 residuals obtained by subtracting the unit means from the observations (Kitanidis 1994). Next the attribute is mapped using universal kriging (Cressie 1991).



**Figure 1.-Nine simulated realities. Letter indicates influence from the artificial soil map: A=large, B=moderate, C=none. Number indicates degree of spatial autocorrelation in the added residual: 1=large, 2=moderate, 3=none.**

In figure 2 the results of the mapping are given for a selection of three out of the nine simulated realities. Note that these are the prediction maps, and that the corresponding maps of prediction error standard deviations are not given here. The DMSV maps necessarily follow the delineations of the soil map, which is quite all right for map A3 but less so for map B2 and definitely inappropriate in case of map C1. Conversely, the CMSV is suitable for mapping C1, but it is not appropriate for mapping B2 and even less so for mapping A3. The most important observation from figure 2 is that the MMSV is indeed capable of an adequate mapping in all three cases. It is interesting to observe that the MMSV mimics the DMSV in case of a 'discrete reality' and the CMSV in case of a 'continuous' reality.

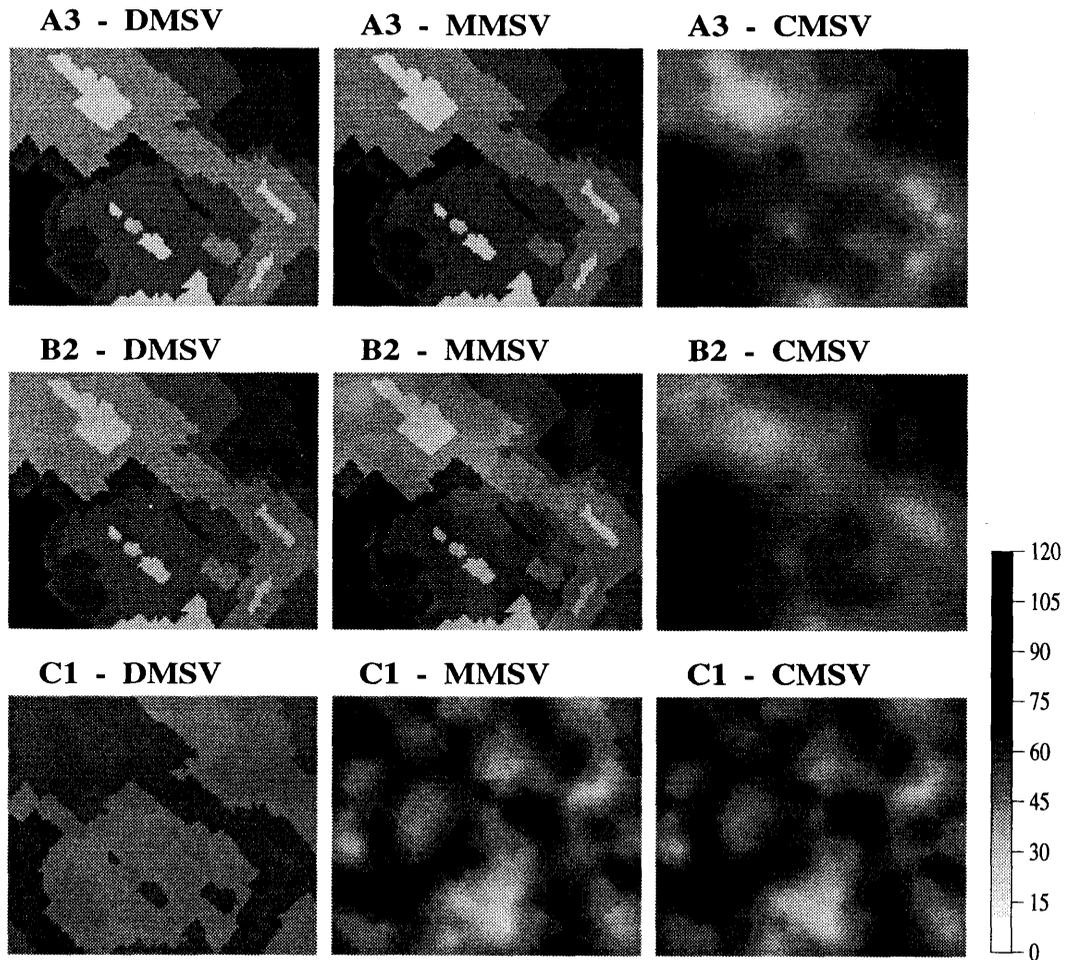


Figure 2.-Mapping three simulated realities (maps A3, B2 and C1) from 200 observations using the DMSV, MMSV and CMSV.

### Validation

In order to evaluate the three prediction methods the mean error (ME), root mean square error (RMSE) and standardized root mean square error (SRMSE) were computed for each of the 9 realities. These statistics were computed from all remaining points in the map. The results are given in figure 3.

The mean error is in all cases quite small. This is not surprising, because unbiasedness conditions are included in all three mapping procedures. Differences between the three mapping procedures are also negligible.

The SRMSE values are on average somewhat larger than one, particularly when the CMSV is applied to situations in which the soil map influence is dominant. This may be caused by forcing the CMSV upon a non-stationary

MMSV or DMSV reality, and perhaps also because the kriging variance does not include the uncertainty in estimating the variogram (Christensen 1991).

Most interesting are the RMSE results. In this case we do see meaningful differences between the three mapping procedures. The results demonstrate that the CMSV is inappropriate for the A-maps, whereas the DMSV is inappropriate for the C-maps. An exception is the pure nugget map C3, where all mapping procedures are equally good (bad). The results also confirm that the MMSV is superior for the B-maps. Note that comparison of RMSE values between maps is difficult here because these are affected by differences in spatial autocorrelation.

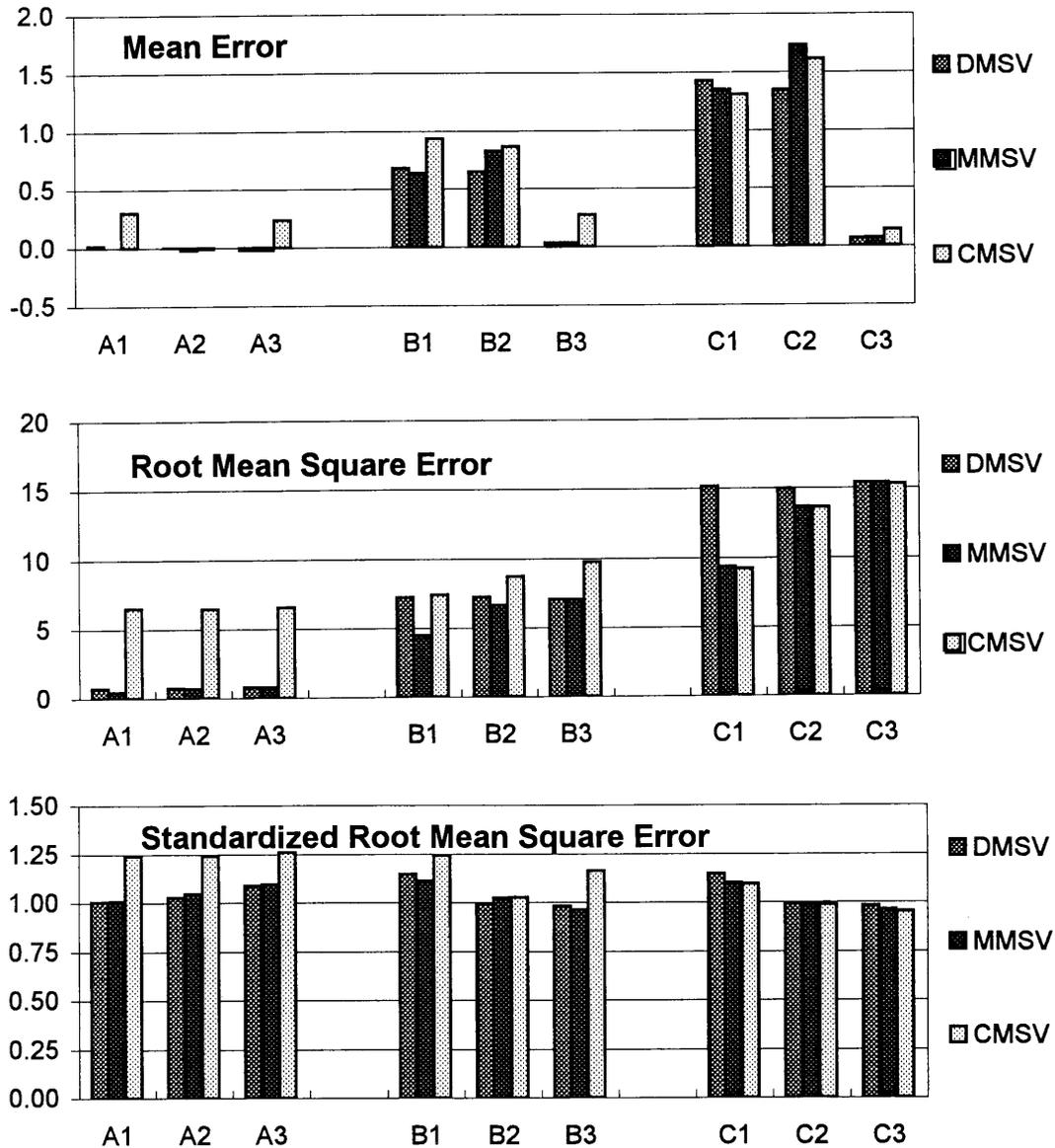


Figure 3.-Validation results for the three prediction methods.

## DISCUSSION AND CONCLUSIONS

The application to simulated 'realities' shows that the MMSV interpolates well on the entire range of spatial variation. In all cases is the MMSV at least as good as the DMSV and the CMSV, and it is superior in situations where there is abrupt and gradual spatial variation.

The simulation exercise also shows that, depending on the situation, the DMSV may perform much worse than the CMSV, and vice versa. This means that if a choice between these two models is to be made (and these are the two models most often used in practice), then it must be taken with care. This may seem obvious, but in practice the choice of model is often dominated by irrelevant factors, such as background and experience of the user. And even when care is taken, it may not always be easy to decide beforehand whether abrupt or gradual spatial variation prevails. Therefore the flexibility of the MMSV demonstrated here is of clear importance, because it implies that by adopting the MMSV one can protect oneself against using the wrong model. It is as if one can leave the choice between abrupt and gradual spatial variation to the MMSV.

The MMSV is especially advantageous in situations where abrupt and gradual spatial variation are both present. Some indication of whether this is the case can be obtained from the intra-class correlation (Webster and Oliver 1990), but more informative is the comparison of the variograms of the original attribute and its residual. A mixed form of spatial variation yields a residual variogram that is substantially lower than the original variogram. Since both variograms can easily be computed one can thus quickly decide whether the MMSV is superior to the CMSV and DMSV for a given situation.

Another approach to handling abrupt and gradual spatial variation both present in the same area is to adopt the CMSV separately per mapping unit (Stein et al. 1988, Voltz and Webster 1990). The main difference with the MMSV is that this approach excludes the presence of spatial autocorrelation across mapping unit boundaries. Thus it is likely to create boundaries even when they are not really there. As mentioned, the MMSV does not suffer from this problem because it mimics the CMSV under such circumstances. We consider it a major advantage of the MMSV that, although it is meant for situations in which gradual and abrupt spatial variation are both present, it will also perform well when spatial variation is exclusively gradual or exclusively abrupt.

## REFERENCES

- Burgess, T.M. and R. Webster. 1980. Optimal interpolation and isarithmic mapping of soil properties. I. The semi-variogram and punctual kriging. *J. Soil Science* 31, 315-331.
- Burrough, P.A. 1993. Soil Variability: a late 20th century view. *Soils and Fertilizers* 56, 529-562.

- Campbell, D.J., D.G. Kinniburgh and P.H.T. Beckett. 1989. The soil solution chemistry of some Oxfordshire soils: temporal and spatial variability. *J. Soil Science* 40, 321-339.
- Christensen, R. 1991. Linear models for multivariate, time series, and spatial data. New York: Springer.
- Cressie, N. 1991. Statistics for spatial data. New York: Wiley.
- Deutsch, C.V. and A.G. Journel. 1992. GSLIB: geostatistical software library and user's guide. New York: Oxford University Press.
- Goovaerts, P. and A.G. Journel. 1995. Integrating soil map information in modelling the spatial variation in continuous soil properties. *European J. Soil Science* 46, 397-414.
- Heuvelink, G.B.M. 1996. Identification of field attribute error under different models of spatial variation. *Int. J. GIS* (in press).
- Heuvelink, G.B.M. and M.F.P. Bierkens. 1992. Combining soil maps with interpolations from point observations to predict quantitative soil properties. *Geoderma* 55, 1-15.
- Kitanidis, P.K. 1994. Generalized covariance functions in estimation. *Mathematical Geology* 25, 525-540.
- Nettleton, W.D., B.R. Brasher and G. Borst. 1991. The taxadjunct problem. *Soil Sci. Soc. Am. J.* 55, 421-427.
- Nortcliff, S. 1978. Soil variability and reconnaissance soil mapping: a statistical study in Norfolk. *J. Soil Science* 29, 403-418.
- Rogowski, A.S. and J.K. Wolf. 1994. Incorporating variability into soil map unit delineations. *Soil Sci. Soc. Am. J.* 58, 403-418.
- Stein, A., M. Hoogerwerf and J. Bouma. 1988. Use of soil-map delineations to improve (co-) kriging of point data on moisture deficits. *Geoderma* 43, 163-177.
- Voltz, M. and R. Webster. 1990. A comparison of kriging, cubic splines and classification for predicting soil properties from sample information. *J. Soil Science* 31, 505-524.
- Webster, R. 1994. The development of pedometrics. *Geoderma* 62, 1-15.
- Webster, R. and H.E. De La Cuanalo. 1975. Soil transect correlograms of North Oxfordshire and their interpretation. *J. Soil Science* 26, 176-194.
- Webster, R. and M.A. Oliver. 1990. Statistical methods in soil and land resource survey. Oxford: Oxford University Press.

### **BIOGRAPHICAL SKETCH**

Gerard B.M. Heuvelink is a geostatistician with the Landscape and Environmental Research Group, Faculty of Environmental Sciences, University of Amsterdam. He holds an M.S. in Applied Mathematics from Twente Technical University and a Ph.D. in Geography from Utrecht University.

Johan A. Huisman is a graduate student in Physical Geography at the Faculty of Environmental Sciences, University of Amsterdam.

# Stratified Kriging Using Vague Transition Zones

G. BOUCNEAU<sup>1</sup>, M. VAN MEIRVENNE<sup>1</sup>,  
O. THAS<sup>2</sup> and G. HOFMAN<sup>1</sup>

---

**Abstract.**— Stratification followed by within-stratum interpolation is a widespread spatial prediction procedure in soil inventories. Conventional stratification assumes that strata are delimited by accurate, crisp boundaries. However, map delineations are not equally accurate. Inaccuracy can result from the inherent gradual nature of the soil property, or errors made during the mapping process. This paper presents a model to describe the vagueness of stratum boundaries and proposes a modified within-stratum kriging algorithm to account for the uncertainty due to this vagueness. The procedure was used to map topsoil sand in Belgium.

---

## INTRODUCTION

Spatial prediction of soil properties is unavoidable when creating a soil information system (SIS). However, every prediction procedure has its shortcomings. Choropleth soil maps assume abrupt changes at their boundaries and interpolation techniques, like kriging, take the condition that variation is gradual and stationary. However, in reality heterogeneous areas often contain both gradual and discontinuous changes of soil properties. Therefore, a combination of both procedures has been suggested to combine the best of both. Stratification of a study area based on soil map delineations, followed by within-stratum interpolation, has been the most widespread combined approach. Stein *et al.* (1988) calculated a variogram for each soil type, Voltz & Webster (1990) computed a pooled variogram and Van Meirvenne *et al.* (1994) standardised the pooled variogram. Other approaches have been the formulation of a mixed model of spatial variation (Heuvelink & Bierkens 1992), or more recently the introduction of a non-parametric indicator approach combined with qualitative information provided by soil or geological maps (Goovaerts & Journel 1995).

---

<sup>1</sup> Soil scientist, Dept. of Soil Management and Soil Care, University of Gent, Coupure 653, 9000 Gent, Belgium.

<sup>2</sup> Statistician, Dept. of Applied Mathematics, Biometrics and Process Control, University of Gent, Coupure 653, 9000 Gent, Belgium.

Conventional stratification assumes strata are delimited by accurate, crisp boundaries. Though, map delineations are not equally accurate. Inaccuracy can result from the inherent fuzzy nature of the soil property, the poor field work, the distortion or shrinkage of the original paper base map or through poor quality digitisation. Attempts have been made to characterise the accuracy of soil map boundaries. Mark & Csillag (1989) conceptually modified the cartographic epsilon model to describe the uncertainty of class membership near a map boundary. However, they do not elaborate on this in a mathematical way. Burrough (1989) mapped the transition between soil strata based on concepts of the fuzzy set theory. Though, he did not make any link with spatial interpolation.

This paper aims to model the vagueness of stratum boundaries in a mathematical way and to modify the within-stratum kriging algorithm to account for the uncertainty due to this vagueness.

## THEORY

### The Vagueness of Stratum Boundaries

The vagueness of a stratum boundary is described by examining point observations of the considered soil property perpendicular to that delineation (figure 1). The shortest distance between observation  $i$  of property  $y$  in stratum  $j$  and the map delineation between strata  $A$  and  $B$  is  $x_{ij}$ . To avoid artefacts at complex combinations of boundaries, each observation should be used only once,

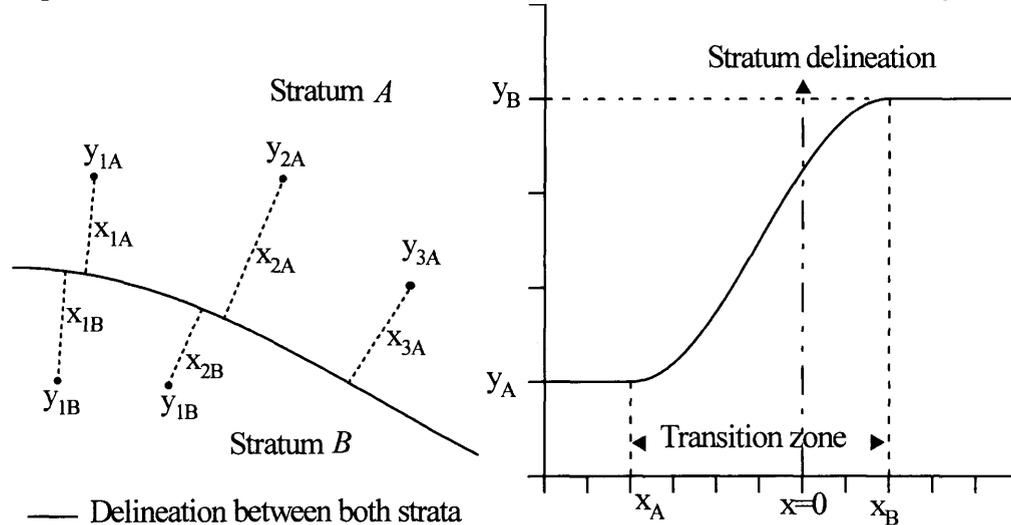


Figure 1.— Examining the vagueness of a stratum boundary using point observations (left), the sigmoidal cosine function describing the vagueness of a stratum boundary (right).

i.e. in respect to the nearest delineation. Then,  $y_{ij}$  is plotted against  $x_{ij}$ . Distances to observations located in stratum  $A$  are given a negative sign, those within stratum  $B$  a positive one. The position of the stratum delineation is at  $x = 0$ . The following sigmoid cosine function can be fitted to model the behaviour of  $y$  at both sides of the boundary (figure 2) :

$$\begin{aligned}
 \text{for } x_{ij} < x_A : \quad & y_{ij} = y_A \\
 \text{for } x_A \leq x_{ij} \leq x_B : \quad & y_{ij} = y_A + (y_B - y_A) \left( 1 - \cos^2 \left( \left( \frac{x_{ij} - x_A}{x_B - x_A} \right) \frac{\pi}{2} \right) \right) \\
 \text{for } x_{ij} > x_B : \quad & y_{ij} = y_B
 \end{aligned} \tag{1}$$

with  $y_A$  and  $y_B$  the typical value of the soil property near the centre of stratum  $A$  respectively  $B$  and  $|x_A| + |x_B|$  the width of the transition zone between both strata. This function was chosen since it allows to model the behaviour of a soil property near a soil boundary asymmetrically in respect to the position of the boundary.

### Within-stratum Variograms

For each stratum the within-stratum variogram was calculated. Observations located within narrow transition zones of inaccurately mapped boundaries were excluded from variogram calculations. For inaccurately delimited strata this resulted in a drastic decrease of the semivariance.

### Prediction With Vague Boundaries

In conventional stratification adjacent strata are delimited by an infinitely thin line and every location belongs to only one stratum. Within-stratum kriging will then result in only one kriging estimate, and one associated variance, for every location of that stratum. In our procedure, boundaries represent a transition zone,  $|x_A| + |x_B|$  wide. Within this zone two kriging procedures,  $K = a$  and  $K = b$ , based on the variograms and data of adjacent strata  $A$  respectively  $B$ , provide two prediction estimates  $\hat{y}_a$  and  $\hat{y}_b$  and two associated estimates of the variances  $s_a^2$  and  $s_b^2$ . In the transition zone and close to stratum  $A$  the structure of the variability will be more similar to the variogram of stratum  $A$  than of stratum  $B$  and the validity of the kriging procedure  $K = a$  is expected to be higher than the validity of the kriging procedure  $K = b$ . The reverse holds for a location within the transition zone and close to stratum  $B$ . The probabilities  $P(K = a)$  and  $P(K = b)$  express the validity of kriging procedures  $K = a$  respectively  $K = b$ . These probabilities are related to the behaviour of the examined soil property in the transition zone. So Eq. 1 is modified as follows:

$$\begin{aligned}
\text{for } x < x_A : & \quad P(K = a) = 1 \text{ and } P(K = b) = 0 \\
\text{for } x_A \leq x \leq x_B : & \quad P(K = a) = 1 - \cos^2\left(\left(\frac{x - x_B}{x_A - x_B}\right)\frac{\pi}{2}\right) \text{ and} \quad [2] \\
& \quad P(K = b) = \cos^2\left(\left(\frac{x - x_B}{x_A - x_B}\right)\frac{\pi}{2}\right) \\
\text{for } x > x_B : & \quad P(K = a) = 0 \text{ and } P(K = b) = 1
\end{aligned}$$

Outside the transition zone and inside stratum  $A$  is  $P(K = a) = 1$  and inside stratum  $B$  is  $P(K = a) = 0$ . Inside the transition zone  $P(K = a)$  decreases from 1 close to stratum  $A$  to 0 close to stratum  $B$ . The reverse holds for  $P(K = b)$ . Estimators of the unconditional expected value  $\hat{E}(Y)$  and its variance  $S^2$  are given by (see appendix for the derivations) :

$$\hat{E}(Y) = Y_a P(K = a) + Y_b P(K = b) \quad [3]$$

$$S^2 = P(K = a)P(K = b)(Y_a - Y_b)^2 - P(K = a)P(K = b)(S_a - S_b)^2 + S_a^2 P(K = a) + S_b^2 P(K = b) \quad [4]$$

To illustrate the influence of  $P(K = a)$  and  $P(K = b)$  on  $S^2$ , three situations were distinguished :

1.  $Y_a$  is twice as large as  $Y_b$  and  $S_a^2$  and  $S_b^2$  are equal. This results in a quadratic curve path of the variance along the transition zone (figure 2a) with a maximum at  $P(K = a) = P(K = b) = 0.5$ . Here  $S^2$  is 3.25 times  $S_a^2$ .
2.  $Y_a$  and  $Y_b$  are equal and  $S_a^2$  is twice as large as  $S_b^2$ . This results in a linearly increasing estimation variances between  $S_a^2$  and  $S_b^2$  (figure 2b).
3. In practice, mostly both the prediction estimates and their variances will differ. In figure 2c both  $Y_a$  is twice as large as  $Y_b$  and  $S_a^2$  is twice as large as  $S_b^2$ . The result is a combination of both preceding situations, with a maximum of  $S^2$  being 3.71 times as large as  $S_a^2$ . It is clear that the difference in prediction estimates dominates the difference in variances.

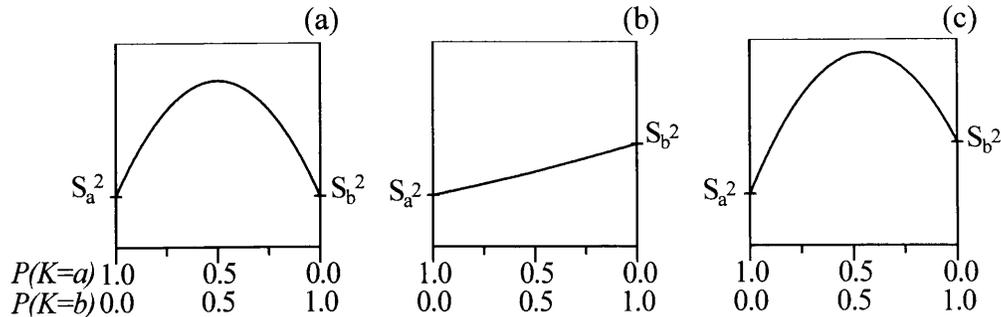


Figure 2.— The prediction variance in the transition zone : (a):  $Y_a = 2 Y_b$  and  $S_a^2 = S_b^2$  ; (b):  $Y_a = Y_b$  and  $S_a^2 = 2 S_b^2$  ; (c):  $Y_a = 2 Y_b$  and  $S_a^2 = 2 S_b^2$ .

It makes sense that the prediction variance inflates at the position of a retained map boundary since this boundary itself represents uncertain information. If one chooses to make use of this information, its uncertainty should be added to the prediction variance.

## RESULTS

The procedure was tested on the top soil sand fraction of West-Flanders, Belgium. More details can be found in Boucneau *et al.* (1996). Stratum boundaries were characterised using Eq. 1. Figure 3 presents an accurately mapped boundary between two strata *B* and *C* (Strata are indicated in figure 4a), having a narrow transition zone of 500 m (left), and an inaccurately mapped boundary between two strata *A* and *C*, having a wide transition zone of 1500 m (right).

A prediction estimate of the topsoil sand and its associated variance was produced based on Eqs. 3 and 4. Figure 4 gives a detail of the map results and allows to compare the proposed procedure with the conventional stratification procedure (i.e. without a transition zone). Three strata can be distinguished: two sandy strata in the upper left corner (stratum *A*) and the lower right corner (stratum *B*) and a stratum with low sand content (stratum *C*) between both. Within the wide transition zone, between the strata *A* and *C*, the proposed procedure recognises the prediction estimates are uncertain by attributing them a high prediction variance (figure 4c). Between strata *B* and *C* the boundary is accurately mapped. Here the transition zone is narrow and high prediction variances are absent.

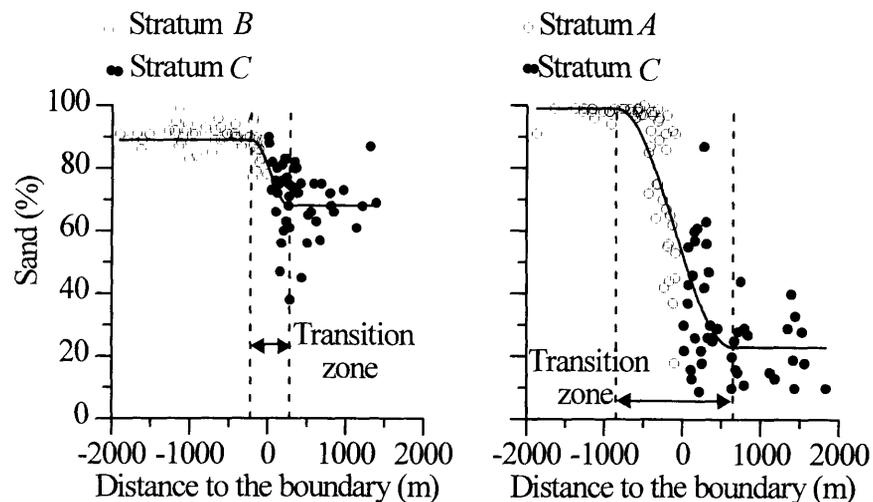


Figure 3. – Narrow (left) and wide (right) transition zones between two strata.

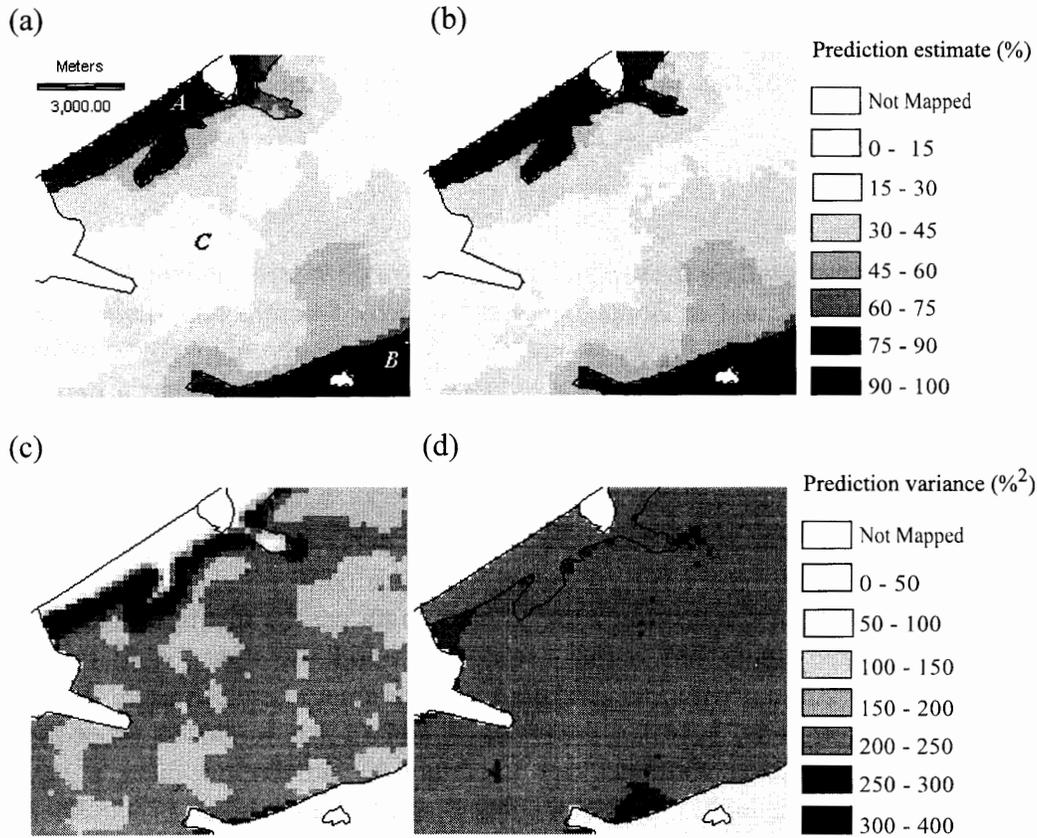


Figure 4.— Detail of the topsoil sand map: the prediction estimate based on the proposed (a) and the conventional procedure (b); the prediction variance based on the proposed (c) and the conventional procedure (d).

Only observations outside the transition zone are used for computing the variogram (figure 5). As a result, the prediction variance in strata *A* and *C* and outside the transition zones, is considerably lower in the proposed procedure than in the conventional procedure (figure 4c and 4d).

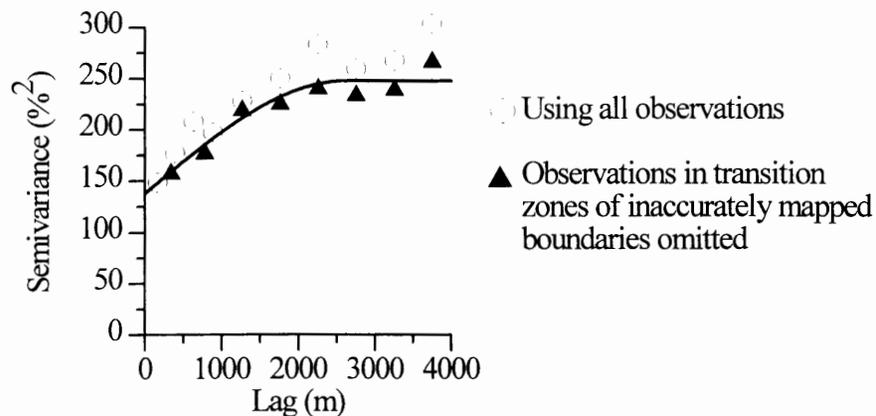


Figure 5.— The within-stratum semivariogram of stratum *C*.

## CONCLUSIONS

In conventional stratification the study area is sharply partitioned in several strata delimited by a thin line. As a result, the within-stratum variogram is based on all data within each stratum. However, soil attribute values of locations situated near the border of a stratum often differ from the typical value of that attribute. These observations can increase the semivariance inside a stratum considerably. In this way the variability inside the transition zone influences the estimate of the variance within the entire stratum. Our procedure overcomes this problem by characterising the transition zone between strata. Only observations outside the transition zone are used for computing the variogram. For each location in a transition zone the probabilities  $P(K = a)$  and  $P(K = b)$  are quantified and incorporated in the prediction estimate and prediction variance. In this way, the variance due to the vagueness of stratum boundaries is concentrated near these boundaries.

## REFERENCES

- Boucneau, G., Van Meirvenne, M., Thas, O. & Hofman G. 1996. Stratified kriging using selected soil map boundaries modelled as vague transition zones. *Submitted to European Journal of Soil Science*.
- Burrough, P.A. 1989. Fuzzy mathematical methods for soil survey and land evaluation. *Journal of Soil Science*, **40**: 477-492.
- Goovaerts, P. & Journel, A.G. 1995. Integrating soil map information in modelling the spatial variation of continuous soil properties. *European Journal of Soil Science*, **46**: 397-414.
- Heuvelink, G.B.M. & Bierkens, M.F.P. 1992. Combining soil maps with interpolations from point observations to predict quantitative soil properties. *Geoderma*, **55**: 445-468.
- Mark, D.M. & Csillag, F. 1989. The nature of boundaries on "Area-Class" maps. *Cartographica*, **26**: 65-78.
- Stein, A., Hoogerwerf, M. & Bouma, J. 1988. Use of soil-map delineations to improve (co-)kriging of point data on moisture deficits. *Geoderma*, **43**: 163-177.
- Van Meirvenne, M., Scheldeman, K., Baert, G. & Hofman, G. 1994. Quantification of soil textural fractions of Bas-Zaïre using soil map polygons and/or point observations. *Geoderma*, **62**: 69-82.
- Voltz, M. & Webster, R. 1990. A comparison of kriging, cubic splines and classification for predicting soil properties from sample information. *Journal of Soil Science* **41**: 473-490.

## APPENDIX

Prediction estimates are obtained based on two kriging procedures  $K=a$  and  $K=b$  which are based on variograms of stratum  $A$  respectively  $B$ . The expected value of the examined variable  $Y$  if the kriging procedure  $K=k$  is valid is  $y_k = E(Y|K=k)$ . The unconditional expected value  $y$  is then obtained by summation over the set of procedures  $K=\{a,b\}$ :

$$y = E(Y) = \sum_{k \in K} E(Y|K=k) P(K=k) = y_a P(K=a) + y_b P(K=b)$$

The variance  $\sigma^2$  is defined by:  $\sigma^2 = Var(Y) = E(Y^2) - \{E(Y)\}^2$ . Working out each term:

$$E(Y)^2 = y_a^2 P^2(K=a) + y_b^2 P^2(K=b) + 2y_a y_b P(K=a)P(K=b)$$

$$E(Y^2) = \sum_{k \in K} E(Y_k^2|K=k) P(K=k) = (y_a^2 + \sigma_a^2) P(K=a) + (y_b^2 + \sigma_b^2) P(K=b)$$

Since:

$$Var(Y_k|K=k) = E(Y_k^2|K=k) - (E(Y_k|K=k))^2$$

$$E(Y_k^2|K=k) = Var(Y_k|K=k) + (E(Y_k|K=k))^2 = y_k^2 + \sigma_k^2$$

And substituting them, the variance  $\sigma^2$  is given by:

$$\begin{aligned} \sigma^2 &= (y_a^2 + \sigma_a^2) P(K=a) + (y_b^2 + \sigma_b^2) P(K=b) - y_a^2 P^2(K=a) - \\ &\quad y_b^2 P^2(K=b) - 2y_a y_b P(K=a)P(K=b) \\ &= y_a^2 P(K=a)(1 - P(K=a)) + y_b^2 P(K=b)(1 - P(K=b)) - \\ &\quad 2y_a y_b P(K=a)P(K=b) + \sigma_a^2 P(K=a) + \sigma_b^2 P(K=b) \\ &= P(K=a)P(K=b)(y_a - y_b)^2 + \sigma_a^2 P(K=a) + \sigma_b^2 P(K=b) \end{aligned}$$

The unbiased estimator for the unconditional mean is:

$$\hat{E}(Y) = Y_a P(K=a) + Y_b P(K=b)$$

$$\text{Since: } y = y_a P(K=a) + y_b P(K=b)$$

The unbiased estimator for the unconditional variance is:

$$\begin{aligned} S^2 &= P(K=a)P(K=b)(Y_a - Y_b)^2 - P(K=a)P(K=b)(S_a - S_b)^2 + \\ &\quad S_a^2 P(K=a) + S_b^2 P(K=b) \end{aligned}$$

$$\text{Since: } \sigma^2 = P(K=a)P(K=b)(y_a - y_b)^2 + \sigma_a^2 P(K=a) + \sigma_b^2 P(K=b)$$

## BIOGRAPHICAL SKETCH

Geert Boucneau prepares a Ph.D. in Soil Information Systems. Marc Van Meirvenne is professor Soil Information Systems and Spatial Data Analysis. Olivier Thas is active in the domain of Environmental Statistics. Georges Hofman is professor Soil Science and Soil Fertility.

# The Effect of Uncertain Locations on Disease Cluster Statistics

Geoffrey M. Jacquez<sup>1</sup> and Lance A. Waller<sup>2</sup>

---

**Abstract.** Disease clusters are space-time aggregations of adverse health events such as birth defects and cancer. Because of incomplete information the exact locations of health events are often unknown, but to date the effect of uncertain locations on disease cluster statistics has not been assessed. This research demonstrates that uncertain locations, if not accounted for, lead to test statistics, null distributions and p-values that differ substantially from those based on actual locations, and increases the probability of type II error (false negatives). Further research is needed to develop statistical methods and software to account for uncertain locations in disease cluster investigations.

---

## INTRODUCTION

Public health agencies use cluster statistics to assess whether an alleged aggregation of disease cases is statistically unusual. The health agency may choose to stop a cluster investigation when the alleged cluster appears to be due to chance, otherwise a more detailed and expensive investigation may be warranted (Centers for Disease Control 1990). Point-based statistics assume exact geographic locations and are often used when health events are rare and the number of observations is small. They are coming into increasing use as spatially referenced health data become commonly available (Openshaw 1991), and models of spatial point processes (Diggle 1993, Lawson and Waller 1996), become increasingly sophisticated. However, the assumption of exact locations upon which point-based statistics are founded is often invalid. Does the violation of this assumption make a difference?

### Sources Of Location Uncertainty

Point based statistics assume data of the form  $(x, y, t)$  where  $x, y$  is the geographic coordinate (such as place of residence) and  $t$  is the date of onset, diagnosis or death. More often than not this location model is inappropriate because locations of the health events are uncertain. This uncertainty has several sources. It emerges when centers of areas such as zip-code zones or census tracts are used instead of exact place of residence. Uncertainty also arises when data are gridded (as, for example, in raster-based GIS) and the coordinates of the nearest grid node are used instead of exact locations. Conceptually, locations are almost always uncertain because humans are mobile rather than sessile and because

<sup>1</sup>President, BioMedware, Ann Arbor, MI.

<sup>2</sup>Assistant Professor, University of Minnesota, Minneapolis

health events and their causative exposures may occur anywhere within a persons activity space. Tobler *et al.* (1995) observed that in modern society a persons daily activity space is approximately 15 km and varies widely. Exact locations do not represent this location uncertainty.

There are many examples of uncertain locations in the literature. Cuzick and Edwards (1990) used the centers of postal code zones to represent place of residence in their study of childhood leukemia in North Humberside. Waller *et al.* (1995) mapped locations of cases of childhood leukemia in Sweden at parish centroids, and evaluated proximity to nuclear power plants using several methods. Location uncertainty arises in these two examples because area centroids are used instead of exact locations. In a study notable for its spatial resolution, Lawson and Williams (1994) used place of residence to assess a possible cluster of deaths from respiratory cancer near a smelter in Armadale, Scotland. Indoor air quality is known to differ substantially from that outdoors, and an unknown proportion of each individuals exposure to smelter fumes presumably occurred outside of the home. In this example location uncertainty arises because place of residence is used to represent exposures that occurred throughout each persons activity space.

We have not evaluated whether or not uncertain locations had an effect, if any, on these studies. Rather, our objective in citing these examples is to give readers a feeling for the ubiquity of the location uncertainty problem. In practice geographic locations are, to a greater or lesser extent, inherently uncertain because they are used as proxy measures of exposure. If we had more information we could explore dose-response relationships using controlled studies. When such detailed knowledge is lacking we instead use spatial relationships among cases, and their proximity to putative sources of hazard, as an uncertain measure of exposure. However, almost all point-based cluster statistics assume the spatial coordinates of health events are precise. What is the impact, if any, of using statistics that assume exact locations when the locations are actually uncertain?

## METHODS

Perhaps the most frequently used paradigm for dealing with uncertain locations is the centroid model. This model assigns all cases occurring within an area the location of that area's centroid. It was used explicitly in the studies in Humberside and Sweden and is implied in the Armadale study. It obtains whenever data are gridded and is common in environmental epidemiology where insufficient knowledge of exposures and confidentiality concerns preclude exact knowledge of health event locations. Using a simulated AIDS epidemic in Michigan, we explored the performance of three disease cluster statistics at three levels of spatial resolution. We compared the test statistics, null distributions, p-values and statistical power under the centroid model to those obtained when actual locations were used.

## Cluster Statistics

Three statistics for space-time clustering were compared, Mantel (1967), Knox (1964) and  $k$ -NN (Jacquez 1996a). These commonly used tests were chosen because they are based on different spatial proximity measures; geographic distance, adjacencies and nearest neighbor relationships, respectively (Jacquez 1996b). We choose tests based on several different proximity measures so that we could generalize the results to a wider range of other cluster tests.

Mantel's test is based on regression of the space distance between pairs of cases on the time distance between case pairs. This study used the reciprocal transformation suggested by Mantel. The Knox test employs space and time critical distances to classify pairs of cases as either near or far in space and time. For example, two cases are considered near to one another when the distance between them is less than the space critical distance. In this study we selected critical values to maximize the Knox test statistic, and used these critical values for all of the simulations. While choosing critical values in this fashion precludes statistical hypothesis testing, it was deemed appropriate for the purpose of the exploratory analyses in this study. The  $k$ -NN statistic is the intersection of the space and time nearest neighbor relationships and is described by Jacquez (1996a). All cluster statistics were calculated using the Stat! software (Jacquez 1994).

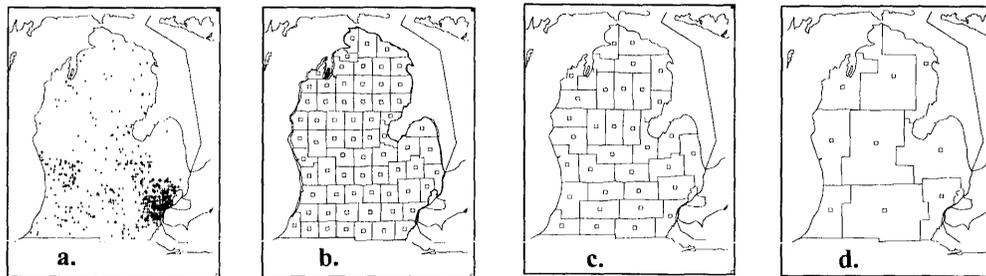


Figure 1. Boundaries corresponding to four levels of geographic uncertainty.

## Geography

Four levels of geographic uncertainty were evaluated: none, county, bicounty and region (Figure 1). Places of residence of the at-risk population are shown in 1a by the '+' symbol and were sampled based on population density distributions of the global demography project (Tobler, Deichmann *et al.* 1995). Centroids and areas are shown for counties (1b), bicounties (1c) and regions (1d). Area centroids are shown as squares.

## HIV Simulation

A simulated epidemic of AIDS in Michigan was used to generate space-time distributions of cases, and is based on the model of Jacquez *et al.* (1994). This HIV

model was chosen because it describes the dynamics of the epidemic and because the transmission parameters have been estimated. The model was simplified for this research and, while the simplified version cannot be said to represent HIV in human populations, it is a contagious process and should result in space-time clusters. If the cluster statistics perform poorly for this model we have little reason to expect them to perform well for diseases such as environmentally caused cancer, for which space-time clustering should be substantially weaker.

The modeled disease has 4 stages: healthy, primary infection, seropositive, and deceased. The vector of transmission probabilities was (0.0, 0.3, 0.001, 0.0). Each element of this vector describes the probability, for a healthy individual, of infection per contact with an individual in disease stage  $i$ . The underlying population consists of 500 individuals with a heterogeneous geographic distribution (Figure 1a). Two of these were selected as seeds to start the epidemic. At each time step each individual contacts 1 of any of 4 nearest neighbors, and a transmission event can occur when the contacted individual is infectious. Transition from healthy to primary infection does not occur in the absence of a transmission event. Probabilities describe transition from one disease stage to another. In each round there is a 5% probability that individuals in the primary infection stage will become seropositive, while, on average, 1% of seropositives enter the terminal stage.

The simulation is a three-step process; first transmission events (which cause healthy individuals to transit to primary infection) are evaluated. Next, stage transitions are determined using transition probabilities. Finally, after each iteration is complete the space-time location of each new seropositive is recorded and the simulation halts when 50 individuals become seropositive. The seropositive stage thus was used for the reporting of disease clusters. The model is stochastic and uses a uniform random number generator to evaluate the discrete transitions for each individual in the population. It is realistic because seropositives are monitored, and because stochasticity is introduced through probabilistic processes of contact, transmission and state-to-state transition.

### **Simulation Protocol**

The HIV simulation model was run 252 times using the actual locations shown in Figure 1a, and new locations of the 2 initial infectives were chosen for each run. Each of these runs is a distinct realization of the AIDS epidemic and is called an 'AIDS simulation'. Each AIDS simulation produced a list of coordinate pairs describing the actual locations of 50 seropositives. Centroid locations for the county, bicounty and regional aggregation levels were assigned as appropriate (Figure 1). For example, at the county level a case occurring in county  $i$  was assigned the coordinates of that county's centroid. The actual test statistic ( $\Gamma_A$ ) and reference (null) distribution ( $g_A$ ) were calculated using the actual locations, and usually are not observable in field situations when centroids are used. The

centroid test statistic ( $\Gamma_C$ ) and its null distribution ( $g_C$ ) were then calculated using the centroid locations, and correspond to statistics calculated in studies where centroid locations are used. The  $g_A$  and  $g_C$  distributions used 249 randomization runs to generate reference distributions expected under the null hypothesis of no association between a case's location and the time at which the case became seropositive. This was accomplished by reallocating the times the cases became seropositive at random across the case locations. These statistics and distributions were recorded for each AIDS simulation and for each cluster statistic.

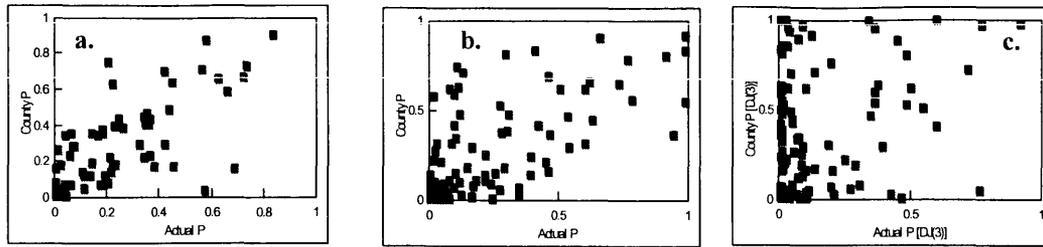
In all, 252 AIDS simulations were run. 249 Monte Carlo randomizations were used to generate each statistical null distribution. 188,244 Monte Carlo randomizations were needed to generate the  $g_A$  distributions (252 AIDS simulation x 3 cluster tests x 249 randomizations) and 564,732 Monte Carlo randomizations were needed to generate the  $g_C$  distributions (252 AIDS simulation x 3 cluster tests x 3 uncertainty levels x 249 randomizations). A grand total of 752,976 Monte Carlo randomizations were conducted using spatial uncertainty modeling software being developed by BioMedware.

## RESULTS

Inquiry focused on four questions. How does uncertainty regarding the spatial locations of health events affect (1) the test statistic, (2) its null distribution, (3) p-values, and (4) statistical power of point-based cluster statistics?

Figure 2 shows the results at the county level. For Mantel's test, test statistics calculated from the centroid locations were consistently lower than those calculated using actual locations, and the upper 95% critical value of the  $g_C$  reference distribution was not highly correlated with the 95% critical value of the  $g_A$  reference distribution calculated from the actual locations. As a result, p-values using the centroid locations were essentially independent of p-values based on the actual locations (Figure 2a). This pattern of results holds for both the Knox test (Figure 2b) and the  $k$ -NN test (Figure 2c). Results, not shown due to space constraints, at the bicounty and regional aggregation levels are as bad or worse and demonstrate, for the HIV model and at the spatial scales considered, that p-values obtained under the centroid location model differ substantially from p-values based on actual locations. This is consistent with the results of Waller and Turnbull (1993), who reported changes in p-values at the census tract and block group aggregation levels for focused cluster tests.

For the 3 cluster statistics considered there is a substantial decrease in statistical power at the county, bicounty and regional aggregation levels. Statistical power was quantified as the proportion of simulation runs for which the p-value was less than or equal to 0.05. Using actual locations the statistical power was 0.592, 0.480 and 0.792 for the Mantel, Knox and  $k$ -NN tests, respectively. At the county level the drop in statistical power ranged from 3.67% (Knox) to 36.62% ( $k$ -NN). At the bicounty level the drop ranged from 6.58% (Knox) to 44.1% ( $k$ -NN). At



**Figure 2. Results at the county level, centroid p-values vs. actual p-values for Mantel (a) Knox (b) and *k*-NN (c) tests.**

the regional level the drop in statistical power ranged from 22.60% (Knox) to 59.83% (*k*-NN).

These results illustrate that p-values from centroid data bear little resemblance to p-values based on actual locations. This causes a loss in statistical power when using centroid locations, and an increase in type II error (false negatives).

### CONCLUSION

As mentioned in the introduction, cluster statistics are one of several criteria used by public health agencies to determine the future course of a cluster investigation. In such situations, p-values from disease cluster tests are used to quantify how unusual an observed aggregation of cases is. A very small p-value suggests the aggregation was unlikely to have been a chance event, and additional effort may be needed to identify environmental exposures that might explain the cluster. While generalization from a single simulation study is difficult, our results suggest that p-values based on uncertain locations are unreliable and should not be used as a quantitative basis for determining the future course of cluster investigations.

Several authors have commented on the apparent lack of statistical power in cluster investigations. This study suggests that actual clusters may be missed because of the loss of statistical power due to location uncertainty.

Even when place of residence is known to the level of the street address, geographic location is often uncertain because health events occur throughout a person's daily activity space, and because exposures (*e.g.* transmission events for infectious diseases, exposure to radiation and carcinogens for cancer, *etc.*) underlying health events occur outside as well as inside the home. This indicates a substantial need for ways of accounting for uncertain locations in general, and for point-based statistics in particular.

There are broader implications because the cluster tests considered are similar to statistics used in other fields. For example, nearest neighbor statistics are used extensively in ecology, and Mantel's test can be expressed as the Pearson product moment correlation which is used extensively in the environmental sciences. All GIS's employ data models that, depending on the GIS operation, may abstract locations to grid nodes or polygon centers, thereby introducing location

uncertainty. This study therefore may have implications for statistical analyses of data managed by Geographic Information Systems.

In conclusion, the software and methods developed in this research allow one to assess, given a specific geography and hypothesized space-time process, the impact of uncertain locations on test statistics, reference distributions, p-values and statistical power. Generalization from the single simulated disease process and geographies considered is problematic. Will the pattern of results hold at finer geographic resolutions? Because of their use in small area studies, it seems particularly important to assess the impact of uncertain locations for census tract and zip-code zone geography. Now that a substantial effect at the county level has been demonstrated, the methods developed in this preliminary study can be used to address this question and others like it.

### ACKNOWLEDGMENTS

This research was funded by Small Business Innovation Research grant R43 CA65366-01A1 from the National Cancer Institute. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NCI.

### REFERENCES

- Centers for Disease Control 1990. Guidelines for investigating clusters of health events. Morbidity and Mortality Weekly Report **39**: 1-23.
- Cuzick, J. and R. Edwards 1990. Spatial clustering for inhomogeneous populations. Journal of the Royal Statistical Society **52**: 73-104.
- Diggle, P. J. 1993. Point process modelling in environmental epidemiology. Statistics for the environment. V. Barnett and K. F. Turkman. New York, John Wiley & Sons Ltd. pp. 89-110.
- Jacquez, G. M. 1994. Stat! statistical software for the clustering of health events. Ann Arbor, MI, BioMedware.
- Jacquez, G. M. 1996a. Disease cluster statistics for imprecise space-time locations. Statistics in Medicine **15**: In press.
- Jacquez, G. M. 1996b. A  $k$ -nearest neighbor test for space-time interaction. Statistics in Medicine In press.
- Jacquez, J. A., J. S. Koopman, et al. 1994. Role of the primary infection in epidemics of HIV infection in gay cohorts. Journal of Acquired Immune Deficiency Syndromes **7**: 1169-1184.
- Knox, G. 1964. The detection of space-time interactions. Applied Statistics **13**: 25-29.
- Lawson, A. B. and L. A. Waller 1996. A review of point pattern methods for spatial modelling of events around sources of pollution. Environmetrics In press.

- Lawson, A. B. and F. L. R. Williams 1994. Armadale : A case-study in environmental epidemiology. Journal of the Royal Statistical Society **157**: 285-298.
- Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Research **27**: 201-218.
- Openshaw, S. 1991. A new approach to the detection and validation of cancer clusters : A review of opportunities, progress and problems. Statistics in Medicine. F. Dunstan and J. Pickles. Oxford, Clarendon Press: 49-63.
- Tobler, W., U. Deichmann, et al. 1995. The global demography project. Santa Barbara, National Center for Geographic Information and Analysis.
- Waller, L. A. and B. W. Turnbull. 1993. The effects of scale on tests of disease clustering. Statistics in Medicine **12**: 1869-1884.
- Waller, L. A., B. W. Turnbull, et al. 1995. Detection and assessment of clusters of disease: An application to nuclear power plant facilities and childhood leukemia in Sweden. Statistics in Medicine **14**: 3-16.

### **BIOGRAPHICAL SKETCH**

Geoffrey M. Jacquez is President of BioMedware, Inc. of Ann Arbor, Michigan. He holds an M.S. in resource policy and law from the University of Michigan, and a Ph.D. (1989) in ecology and evolution from SUNY Stony Brook.

Lance Waller is Assistant Professor of Biostatistics in the School of Public Health at the University of Minnesota. He received his Ph. D. in operations research from Cornell University in 1991.

BioMedware was founded in 1990, develops software and statistical methods for the environmental and health sciences, and is the recipient of SBIR and STTR grants from the NIH.

# Reliability of Area Mapping by Delineation in Aerial Photographs

Claus-Peter<sup>1</sup> Gross and Petra Adler<sup>2</sup>

---

**Abstract.** - In a test series under standardised conditions, findings between individual photo-interpreters concerning forest stand boundaries delineation were investigated. It was found that the dispersion of the delineation was dependent on the degree of difficulty in defining a natural border by the interpreter. The evaluation showed that 90% of delineated polygon lines of a „normal“ mapping project under central European forest conditions were found in a 10 meter buffer with a photo scale 1:7000 and required 22 meter buffer in a photo scale 1:20000 to fit 90% of the delineated polygon lines.

---

## INTRODUCTION

Aerial photographs are an important source of actual and detailed area information. For decades the delineation of polygons of homogeneous information from images for forestry or landuse application has been common practice. After geometric correction using some kind of photogrammetric procedure the stand polygons can be integrated as an information layer in a Geographical Information System (GIS).

In dependence of the photogrammetric solution (low budget, advanced photogrammetry), the positional error of the polygon lines can be estimated. However, there is still uncertainty concerning the scope and possible variance of interpretation, because:

- the photogrammetric measurement of lines is less precise than the measurement of point objects.
- forests are natural objects, in general without well defined, clear borders.
- the clearness of borders between different forest stands varies to a wide degree.

Besides the photogrammetric accuracy the interpreter himself is responsible for an unknown uncertainty. It is of great interest for GIS analysis to have an quantitative idea of the influence of this human factor on the spatial data.

---

<sup>1</sup> Forester, specialised in Remote Sensing (Ph.D.), Dept. of Remote Sensing and Landcsapinformationsystem, University of Freiburg, Germany

<sup>2</sup> Student of Forestry and Remote Sensing, University of Freiburg, Germany

## **MATERIAL AND METHODS**

In the study aerial Color-Infrared (CIR) photographs of two different scales were evaluated. The large scale photographs (1:7000) show an intensively managed forest district in Germany (Duvenhorst, 1995). Small scale aerial photographs (1:20000) were taken from a forest region with little management activities in Italy (Gross, 1993). Both photo-flights were taken for use in an intensive forest inventory based on remote sensing.

In a first step the degrees of difficulty to identify the forest borders was estimated in three classes (easy, medium, difficult) using a representative check of all polygon lines in the test sites delineated in both inventory projects.

In a second step the reliability of the forest stand delineation was tested. From the two test sites a representative study area was chosen for the different photo scale. Ten experienced photointerpreters, students and staff members of the Department had to delineate five forest stand boundaries in study area using defined interpretation keys. For the mapping procedure a high precision photogrammetric instrument (ZEISS Planicom P3) was used to exclude the influence of equipment errors.

A particular problem for the evaluation of the test series is the lack of reference data. The borders of forest, as well as of many landscape objects, are „soft“ and no true borderline can be mapped, using either terrestrial means or aerial photographs. To handle this problem adequately and to quantify the result of the test, a special GIS- application was chosen to evaluate the test series:

- All polygon section in which the test persons easily identified the same stand border were cut out of the polygons for further evaluations. A fundamentally different interpretation of a forest stand, e.g. the division of a stand in smaller unites, was not of central interest in this study. These variations can be influenced through training, interpretation guidelines and the experience of the interpreters.
- Starting from the minimal and maximal polygon line of the entire bundle, the percentage of the total line length in 1 meter (1:7000) and 2 meter (1:20000) buffers was calculated.
- The raw data were put in a new order, starting with the buffers with highest density of polygon lines, independent of their localisation. The results presented in this paper show the average between the results for maximum and minimum polygons as cumulative histogram.

## **RESULTS**

How difficult is it for experienced forest photointerpreters to identify and delineate forest stand boundaries from aerial photographs? Independent of the scale of the photographs it was found, that about 2/3 of all stand limits were



Figure 1: Degree of difficulty to identify forest stand borders in aerial photographs in the scale of 1:7000 (left) and 1:20000 (right).

clearly visible and can be identified without any doubt. Only about 20% of the forest stands were found to be extremely difficult to delimit to neighbouring stands. These borders were found to be more or less fluid (figure 1). Of course, the visibility might change with natural forest conditions and silvicultural concepts. For the situation in central Europe these results were considered quite representative.

The delineation test series showed the following results for the two photo scales:

- In the **large scale** photographs the variation of easy to identify polygon lines was a maximum of 5 meters. More than 80% of all lines were found in a buffer of 2 meter width. The delineation of medium difficult to identify lines varied for 91% in a 11 meter dispersion whereas a buffer width of 27 meter was necessary to fit 90% of all lines that were really difficult to interpret (figure 2).

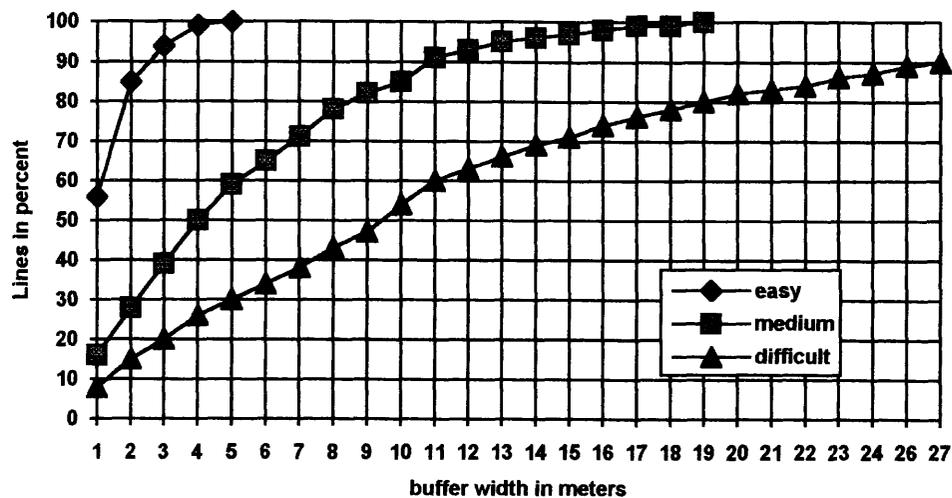


Figure 2: Variation of delineation results in three degrees of identification difficulty (photo scale 1:7000)

- The **small scale** photographs naturally show a higher dispersion. A buffer of 10 meters fits 90% of all lines that were easy to detect. This makes 0.5 millimetres (mm) in the photo scale 1:20000. About 90% of the medium difficult stand borders were found in a 38 meter buffer and 90% of the difficult to identify lines varied in a 46 meter width buffer. To fit all polygon lines a buffer of 80 meter was necessary (Figure 3). The precision in coordinates of the stereomodel is not so different from the large scale photos. The 0.5mm in 1:20000 for easy delineated lines is similar to the 0.4mm in 1:7000. The difficult to interpret lines require 2.3mm (1:20000) and 3.8mm (1:7000) to fit 90% of all lines. It seems that the small scale photographs may have even some advantages for the interpretation because of the better overview.

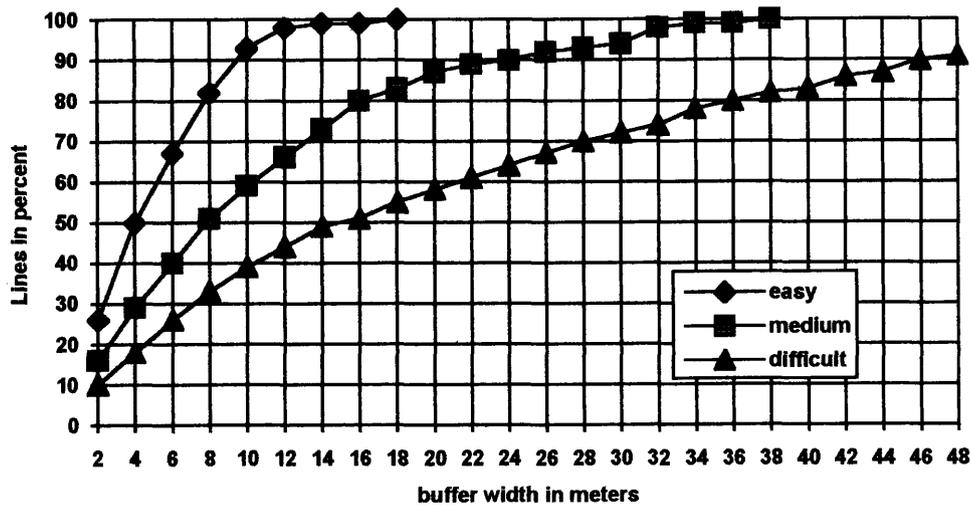


Figure 3: Variation of delineation results in three degrees of identification difficulty (photo scale 1: 20.000)

### CONCLUSION

The central object of this study is the subjectivity of aerial photo interpretation related to spatial accuracy of mapped forest stand border lines. Only very few approaches can be found in the literature to quantify the influence of the human factor (e.g. Congalton 1983, Glemser 1993, Scherrer et al. 1990, Zihlavnik 1991). To identify the correct stand border well worked and defined interpretation keys and sufficient training of the interpreters are preconditions. But even when the interpreters identify the same stand border and are willing to map the same line they have some freedom of action. The resulting variations are strongly related to the clearness of the stand delimitation. The photo scale has mainly an effect on the absolute precision in ground coordinates.

When the delineation variations are transferred to the entire inventory projects by weighting the data with the portion of identification difficulty, then 90% of all polygon lines in the 1:7000 photo flight were in a buffer width of 10 meters, in the 1:20000 photographs a width of 22 meters was required. This means, speaking from the point view of a forester, under normal conditions the accessible precision ranges from less than a tree crown to maximal one adult broadleaf crown.

Nevertheless for some GIS procedures such as the intersection of stand area with qualitative data, the existing uncertainty should be taken into consideration. The authors propose to draw buffers with the same probability around the mapped polygon lines with regard to the difficulty of identification.

## REFERENCES

- Congalton, R. G. 1983: A Quantitative Method to Test for Consistency and Correctness in Photointerpretation. *Photogrammetric Engineering and Remote Sensing* 49 (1), p. 69-74
- Duvenhorst, J. 1995: Photogrammetry in Forestry - Tool for Effective Inventory and Planning. in Fritsch, D.; Hobbie, D.(eds.): *Photogrammetric Week'95*, Wichmann Heidelberg, p. 337-349
- Glemser, M. 1993: Untersuchungen zur objektbezogenen geometrischen Genauigkeit. *Salzburger Geographische Materialien*, Heft 20, p. 97-108
- Gross, C.P. 1993: Regionale Waldinventur zur Erfassung des Waldzustandes mit kleinmaßstäbigen Color-Infrarot Luftbildern. *Dissertation Universität Freiburg*, 126 p.
- Scherrer, H. V.; Gautschi, H.; Hauenstein, P. 1990: Flächendeckende Waldzustandserfassung mit Infrarot-Luftbildern. *Bericht der Eidgenössischen Anstalt für das Forstliche Versuchswesen Nr. 318*, Birmensdorf, 101 p.
- Zihlavnik, S. 1991: Utilization of Photointerpretation in the Determination of Forest Stand Boundaries. *Lesnictvi*, p. 819-829

## BIOGRAPHICAL SKETCH

Claus-Peter Gross is a graduated forester (Ph.D.) After five years practical work in the forest administration, he joined in 1987 the remote sensing team at the Forest Faculty in Freiburg (Prof. Hildebrandt) and is now scientific employée at the Department of Remote Sensing and Landscape-Informationssystem (Head Prof. B. Koch). He is specialised in forest remote sensing applications and photogrammetry.

Petra Adler will finish her studies in forestry at the University of Freiburg 1996. In her studies she put a special emphasis on remote sensing.

# The Effect Of Database Generalization On The Accuracy Of The Viewshed

Peter Fisher<sup>1</sup>

---

**Abstract --** This paper examines the effects of database generalization on the area which is determined to be visible in GIS analysis. Many different methods of generalization are possible, but here, for any cell at the target resolution, elevations are determined from: the arithmetic mean, the maximum, the minimum, and the maximum difference from the mean of the cells within the kernel, and all possible combinations of regular spacings. The resolutions studied are 0.5, 0.33, 0.25, and 0.2 of the original study area. The viewsheds determined over these different resolution DEMs are compared with a number of possible viewsheds derived by generalization of the viewshed over the original DEM. Of those tested the maximum deviation from the mean within kernel provides the best estimate of the pattern and area of the viewshed at all resolutions.

---

## INTRODUCTION

In recent years considerable attention has been paid to two related lines of research. The representation of digital spatial data at multiple resolutions, where either information is stored at multiple resolutions or is generalized from a detailed (large) scale to a generalized (smaller) scale (Buttenfield and McMaster 1991; Muller, Lagrange and Weibel 1995). For specific studies see, for example, Brown et al. (1993), Chang and Tsai (1991), Isaacson and Ripple (1990), Joao (1995) and Painho (1995). Few pieces of work have taken a dataset at one single resolution and examined derived products in alternative generalizations of the data, but see work of McMaster (1987).

The research reported here is in this same vein. It takes the viewshed as it is determined in one DEM at the largest resolution, and uses measurements of the 'quality' or 'accuracy' of the viewshed when it is determined from alternative derived DEMs at smaller resolutions. The purpose is to comment on the changing nature of the viewshed and the performance of the alternative generalization operators defined.

---

<sup>1</sup> Senior Lecturer, University of Leicester, Leicester, United Kingdom

## THE VIEWSHED

The viewshed is one of the standard procedures included with GIS designed for analysis of elevation data. It is intended to distinguish locations which can be seen from a particular viewing location (in-view) from those which are out-of-view. A line-of-sight is drawn from the viewer to the viewed location, and if the elevation rises above the elevation of the line-of-sight at any point along the line-of-sight then the viewed position at the end of the line is determined to be out-of-view. Otherwise it is in-view. Here the method of point-to-point determination across a rectangular lattice is used (Fisher 1993).

## EXPERIMENTAL PROCEDURE

For the research reported here two Ordnance Survey DEMs from the United Kingdom were used: one of the Malvern Hills and one including the southeast part of Dartmoor. Each is a 20 x 20 km tile of the national elevation coverage of 1:50,000 gridded elevation data recording elevations at 50 m intervals. Therefore each is 401 x 401 values. Within each test area 100 random points were selected within an active area which excluded a 60-cell buffer from the edge of the tile. A region 120 x 120 cells was then taken around each point. The 60,60 cell within each area was taken as the view point and the viewshed determined. These are binary products which are referred to below as the *original viewsheds*.

Generalizations were then made of the DEM itself, the original viewshed and the viewpoint, so that viewsheds could be determined. For simplicity the original DEMs are generalized so that pixels in the derived datasets map exactly to the original. The generalizations included 0.5, 0.33, 0.25, and 0.2 times reductions, giving DEM arrays of 60x60, 40x40, 30x30 and 20x20. The generalization procedures are discussed in the next section.

Statistical comparisons include the bias and error of the viewshed in the generalized DEMs was derived from a comparison with the area of the original viewshed, and the kappa coefficient of agreement between the viewshed in the generalized DEMs and a number of alternative generalizations of the original viewshed.

## GENERALIZATION OPERATORS

### Generalizing the DEM

It is easy to envisage many alternative approaches to the generalization of a gridded terrain dataset. The methods used here are entirely based on raster filtering. They do not include any surface modelling methods. Another approach is to resample using one of a number of well known algorithms,

including nearest neighbour, and bilinear interpolation. These methods, however, are generally reserved for the transformation of one gridded dataset to another where the orientation of the grid varies, as well as the size of the grid. Other approaches to generalizing the grid use a more regular geometric approach, and are investigated here.

Two primary types of generalization are used. The first type searches for values in the original DEM on a regular spacing basis, while the others take summary statistics from within a kernel area.

### **Regular Spacing**

The simplest method of generalization is to take a value at a regular spacing,  $m$ , so that in the values in the derived DEM are every  $m$ th values in  $x$  and  $y$  in the original DEM. If  $m = 2$  then there are 4 possible realizations of this process.  $z_d = z_1$ , then  $z_d = z_2$ , then  $z_d = z_3$ , and finally  $z_d = z_4$  (where  $z_1, z_2, z_3$ , and  $z_4$  alternative elevation values, and  $z_d$  is the value in a realization of the generalization process). If  $m = 5$  then there are 25 different possible values of  $z_d$ .

### **Summary Statistics**

#### 1) Mean

Within the kernel area the most obvious summary statistic is the mean.

#### 2) Maximum and Minimum

Two further generalization operators are the maximum and minimum values within kernel.

#### 3) Maximum Deviation from the Mean

Finally, the value of the elevation which is the largest deviation from the mean within the kernel is recorded.

## **Generalizing the Viewshed**

For comparison with derived viewsheds at multiple levels of generalization it is necessary to have a generalization of the original viewshed. It should be recalled that the viewshed treated here is a binary phenomenon, any cell can only be treated as being in-view or out-of-view. Three different strategies are used:

#### 1) All

A cell in the generalized dataset is taken to be in-view if all cells in the kernel are in-view. Otherwise it is out-of-view.

#### 2) Majority

A cell in the generalized dataset is taken to be in-view if the majority of cells in the kernel are in-view. Otherwise it is out-of-view. Where there is an even number of cells in the kernel the majority is taken as  $n/2 + 1$ .

#### 3) Any

A cell in the generalized dataset is taken to be in-view if the any cell in the kernel is in-view. Otherwise the generalized cell is out-of-view.

### **Generalizing the Viewpoint**

Finally, the treatment of the viewpoint is considered. Viewshed algorithms treat the viewpoint very differently. Here, the cell considered to be the viewpoint in any generalization is the generalized cell which contains the viewpoint in the ungeneralized DEM. This approach has the advantage over others that the viewpoint is always at the centre of a cell in any dataset, and so coincides geometrically with a height value in the corresponding DEM. The DEM elevation at the viewpoint is always known as part of the generalized DEM and never inferred. Many possible resizations of the original viewshed correspond to the viewshed at a particular scale, and so results here must be regarded as provisional.

### **Implementation**

The above approach was implemented, together with the viewshed algorithm in Turbo Pascal 7.0. The sampling, generalization by all methods, generation of the viewshed and calculation of summary and comparative statistics of the viewsheds were all integrated. It ran on a Pentium PC which had been checked for faults on the processor. Idrisi v 4.1 was used in support of the analysis to view data at different resolutions, etc. and analysis was done in Excel 5.0 and using bespoke programs.

## **RESULTS**

### **Generalizing the Original Viewshed**

As noted above, three different versions of generalizing the original viewshed were derived. These are based on all, a majority and any of the pixels within the kernel being in view in the original. The results of this analysis can best be seen in Table 1, where the correlation between the area of the original and generalized viewsheds is reported together with the mean bias and error as a result of the generalization for both test areas. It can be seen that the correlation for the majority operator is always closes to 1.0, and that it has much the lowest values of bias and error at all levels of generalization. In short, the area of the viewshed derived by the majority operator is predictable, and is very similar to the original area, even over large resolution changes, while the values as a result of the any and all operators are not so predictable at the larger resolution changes. Furthermore, the values of the majority operator are very close to the values for the original.

**Table 1. -- Bias and Error in generalizations of the Original Viewshed.**

Generalizations		Malvern			Dartmoor		
		Bias	Error	Correl.	Bias	Error	Correl.
0.5 x	Any	3.39	3.78	0.993	2.42	2.70	0.997
	Major	-1.19	1.34	0.997	-0.87	0.97	0.999
	All	-3.60	4.06	0.985	-2.48	2.79	0.993
0.33 x	Any	5.98	6.58	0.983	4.50	4.97	0.993
	Major	0.09	0.43	0.999	0.00	0.24	0.999
	All	-6.21	7.02	0.943	-4.45	5.04	0.972
0.25 x	Any	8.36	9.12	0.971	6.46	7.10	0.988
	Major	-0.56	0.77	0.997	-0.50	0.63	0.998
	All	-7.93	8.99	0.881	-5.96	6.75	0.940
0.2 x	Any	10.54	11.46	0.958	8.30	9.09	0.980
	Major	-0.29	0.70	0.997	-0.26	0.61	0.995
	All	-9.13	10.39	0.803	-6.99	7.92	0.896

### The Area of the Viewshed

When the viewshed is determined over the generalized DEMs, there are a number of ways to compare it with the original viewshed. Two parameters are examined here: the area visible, and the kappa coefficient of agreement. The area of the viewshed at different resolutions is easily determined, and in Table 2 the mean bias and error are given for both test areas at all four resolutions. In addition, the correlation coefficients between the two areas are reported. The “best” generalization is that which yields the smallest bias and error and the largest correlation.

Table 2 is divided by the amount of and methods of generalization. The values reported for regular spacings are the minimum area and the maximum area determined among the 4, 9, 16 and 25 possible. There is considerable variation among these, as is reflected in the error measures which are always the largest and smallest of any other generalization. On the other hand, the correlation coefficients for these are on the whole both poorer than the best of the summary statistic values.

Among the summary statistics of the generalizations, 21 out of the 24 measures of accuracy show the maximum deviation from the mean to give the best performance. The correlation coefficients and mean errors yielded by this method are always the best of those reported. The bias is sometimes smaller for some other method (in two cases the maximum within kernel and once the minimum within kernel).

**Table 2. -- Error measures for the area of the viewshed in generalizations of the viewshed as compared with the area in the original viewshed.**

Generalizations			Malvern			Dartmoor		
			Bias	Error	Correl.	Bias	Error	Correl.
0.5 x	regular	Max	6.17	8.17	0.802	3.60	4.34	0.956
	spacing	Min	1.14	3.32	0.913	0.47	1.82	0.965
	summary	Max	4.04	6.56	0.809	2.18	3.18	0.946
	statistics	Min	4.27	6.63	0.823	<b>2.03</b>	3.12	0.957
		Mean	4.52	6.76	0.830	2.23	2.97	0.972
		MaxDev	<b>3.39</b>	<b>4.02</b>	<b>0.967</b>	2.11	<b>2.58</b>	<b>0.992</b>
0.33 x	regular	Max	11.39	13.68	0.706	7.52	8.75	0.871
	spacing	Min	2.28	6.42	0.725	1.03	3.33	0.897
	summary	Max	6.85	9.34	0.768	4.50	6.05	0.874
	statistics	Min	8.51	11.25	0.688	4.57	6.59	0.880
		Mean	7.87	10.10	0.768	4.74	6.17	0.910
		MaxDev	<b>6.60</b>	<b>8.08</b>	<b>0.884</b>	<b>4.08</b>	<b>4.85</b>	<b>0.979</b>
0.25 x	regular	Max	16.34	18.89	0.600	12.26	14.26	0.712
	spacing	Min	3.34	7.42	0.673	1.76	4.88	0.812
	summary	Max	<b>9.55</b>	12.24	0.658	7.46	9.40	0.784
	statistics	Min	12.61	15.86	0.600	7.25	9.48	0.806
		Mean	11.22	14.01	0.675	7.38	9.12	0.836
		MaxDev	9.65	<b>11.23</b>	<b>0.844</b>	<b>6.29</b>	<b>7.53</b>	<b>0.955</b>
0.2 x	regular	Max	20.75	23.23	0.526	16.60	18.87	0.648
	spacing	Min	4.15	8.08	0.651	2.30	5.51	0.738
	summary	Max	<b>12.45</b>	15.88	0.542	9.78	12.20	0.705
	statistics	Min	15.83	18.96	0.580	10.26	12.62	0.685
		Mean	14.00	17.12	0.607	10.14	12.38	0.706
		MaxDev	12.77	<b>14.60</b>	<b>0.790</b>	<b>8.97</b>	<b>10.57</b>	<b>0.927</b>

The latter of these shows an enormous spread of values which tends to be more typical of the 0.2 generalizations than that yielded by the maximum deviation from the mean. The 0.5 generalizations are generally relatively well correlated as is shown by the correlation coefficients, but this can rapidly deteriorate.

### The Arrangement of the Viewshed

The kappa coefficient of agreement between two sets of data ranges from 0 to 1, where 1 reflects perfect correspondence in the arrangements and 0 disagreement. It has been widely applied to examining the confusion matrix in remote sensing (Congalton et al., 1983). Here the same approach is used. Kappa is determined for 2 x 2 tabulations which, for a particular resolution, compares the visibility of pixels in the generalized binary viewshed, as determined by the majority method, with that determined through the DEM when generalized by one of the methods under test. Summary results are

reported in Table 3 where the maximum, minimum and mean values of the coefficient for the 100 viewpoints in each study area are reported. The maximum and minimum values are reporting the best and the worst agreements between patterns in any situation. The best generalization method should yield the largest value of kappa at any reduction.

The results for regularly spaced samples report the agreements for the maximum and minimum areas in any single generalization test. The maximum viewshed is the best in 22 out of the 24 values reported. In other words, the largest viewshed generated by regular spacing from the kernel area yields the best agreement in the spatial pattern of the viewshed among all other generalizations of the DEM. On the other hand, the minimum area viewshed always yields the worst agreement. This method of generalization provides an envelope of all other values. Indeed, in the 0.2 generalizations of both the Malvern and Dartmoor test areas, regular spacing yielding the minimum visible area gives almost total disagreement with the original viewshed (0.054 and 0.020).

**Table 3. -- Kappa coefficients of agreement between the generalized version of the original viewshed and the viewshed determined over the generalized DEM.**

Generalizations			Malvern			Dartmoor		
			Max	Mean	Min	Max	Mean	Min
0.5 x	regular spacing	Max	0.899	0.744	0.285	0.924	0.818	0.661
		Min	0.850	0.591	0.201	0.845	0.688	0.135
	summary statistics	Max	0.878	0.674	0.231	0.903	0.760	0.371
		Min	<b>0.910</b>	0.661	0.217	0.879	0.769	0.414
		Mean	0.902	0.692	0.222	0.890	0.784	0.411
		MaxDev	0.899	<b>0.725</b>	<b>0.263</b>	<b>0.924</b>	<b>0.802</b>	<b>0.440</b>
0.33 x	regular spacing	Max	0.886	0.694	0.235	0.877	0.772	0.480
		Min	0.805	0.465	0.106	0.796	0.540	0.162
	summary statistics	Max	0.868	0.607	0.182	0.864	0.666	0.201
		Min	0.856	0.562	0.158	0.853	0.675	0.316
		Mean	<b>0.873</b>	0.624	0.193	0.858	0.705	0.278
		MaxDev	0.864	<b>0.646</b>	<b>0.196</b>	<b>0.877</b>	<b>0.728</b>	<b>0.480</b>
0.25 x	regular spacing	Max	0.856	0.601	0.170	0.864	0.688	0.427
		Min	0.748	0.353	0.053	0.695	0.406	0.053
	summary statistics	Max	<b>0.833</b>	0.496	0.106	0.779	0.559	0.126
		Min	0.816	0.447	0.102	0.795	0.555	0.245
		Mean	0.831	0.510	<b>0.133</b>	<b>0.829</b>	0.593	0.180
		MaxDev	0.812	<b>0.528</b>	0.118	0.795	<b>0.615</b>	<b>0.379</b>
0.2 x	regular spacing	Max	0.833	0.555	0.174	0.827	0.634	0.238
		Min	0.711	0.288	0.054	0.613	0.320	0.020
	summary statistics	Max	0.785	0.446	0.082	0.813	0.487	0.047
		Min	0.785	0.393	0.015	0.775	0.484	0.067
		Mean	<b>0.825</b>	0.458	0.109	<b>0.815</b>	0.528	0.059
		MaxDev	0.750	<b>0.469</b>	<b>0.118</b>	0.767	<b>0.533</b>	<b>0.193</b>

In 17 out of 24 summaries, the maximum deviation within kernel yields the best agreement. In 5 cases the mean is best, and the maximum and minimum within kernel are best once each. With 0.2 generalization the worst levels of agreement are again very poor (0.015 in the 0.2 generalization of the Malvern area by minimum within kernel). Both the mean and the best levels of agreement in all situations are, however, quite acceptable.

## CONCLUSION

A number of conclusions can be drawn from the current work. For the viewshed it would appear that the generalization of the binary viewshed to other grid resolutions is best performed by determining those cells at the target resolution with the majority of visible cells. Other immediately apparent methods of generalization do not yield results which match the area and pattern of the original viewshed.

Generalization of the DEM yields many alternative possible viewsheds. Regular spacing of the kernel area yields both the best and worst estimates of the visible area. The result is therefore unpredictable, and although the method is fast and convenient it is not to be recommended as a basis of generalization. Among the statistical summaries of a kernel area, generalization of the DEM by determination of the maximum deviation within the kernel yields the viewshed which best reflects both the area and arrangement of the visible area. The results are all most stable and most frequently the best for this method. It is therefore the method to be recommended (of those tested) in any situation where a DEM requires generalization, and the viewshed is the derived product which is uppermost in the investigator's interest.

Finally it should be noted that in most cases neither the area nor the pattern of the viewshed is badly disrupted by generalization of the DEM, although it can be if an injudicious method of generalization is used.

As stated in the introduction, very little research is reported on propagating the effects of alternative generalizations of spatial databases. Research similar to that reported here can be envisaged for almost any spatial data and derived product. Indeed, it is crucial for very many applications that we know best how to generalize both categorical data (the binary viewshed), and continuous data (the DEM), and possible consequences of generalization. While the results reported here seem quite conclusive, and are very likely to be generalizable to other areas, so long as the viewshed is of interest, there is absolutely no guarantee that either the recommended method of generalization should be the same for all derived products, or that there is not a better generalization operator for the viewshed.

## ACKNOWLEDGEMENTS

I would particularly like to thank Jo Wood for some insightful suggestions in the course of this work.

## REFERENCES

- Brown, D.G., Ling Bain and Walsh, S.J., 1993. Response of a distributed watershed erosion model to variations in the input data aggregation levels. *Computers & Geosciences* 19, 499-509.
- Buttenfield, B.P., and McMaster, R.B. (Editors) 1991. *Map Generalization: Making Rules for Knowledge Representation*. London: Longman.
- Chang, K.-t., and Tsai, B.-w. 1991. The effect of DEM resolution on slope and aspect mapping. *Cartography and Geographic Information Systems* 18, 69-77.
- Congalton, R.G., Oderwald, R.G., and Mead, R.A., 1983. Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques. *Photogrammetric Engineering and Remote Sensing* 49, 79-87.
- Fisher, P.F. 1993. Algorithm and Implementation Uncertainty in Viewshed Analysis. *International Journal of Geographical Information Systems* 7, 331-347.
- Isaacson, D.L., and Ripple, W.J. 1990. Comparison of 7.5-minute and 1-degree digital elevation models. *Photogrammetric Engineering and Remote Sensing* 56, 1523-1527.
- Joao, E.M. 1995. The importance of quantifying the effects of generalization. In *GIS and Generalization: Methodology and Practice*, edited by Muller, J.C., Lagrange, J.P. and Weibel, R. London: Taylor & Francis. pp. 183-193.
- McMaster, R.B. 1987. The geometric properties of numerical simplification. *Geographical Analysis* 19, 330-346.
- Muller, J.C., Lagrange, J.P. and Weibel, R. (Editors) 1995. *GIS and Generalization: Methodology and Practice*. London: Taylor & Francis. 257 p.
- Painho, M. 1995. The effects of generalization on attribute accuracy in natural resource maps. . In *GIS and Generalization: Methodology and Practice*, edited by Muller, J.C., Lagrange, J.P. and Weibel, R. London: Taylor & Francis. pp. 194-206.

# Modeling Slope Stability Uncertainty: A Case Study at the H.J. Andrews Long-Term Ecological Research Site, Oregon

Michelle L. Murillo<sup>1</sup> and Gary J. Hunter<sup>2</sup>

---

**Abstract.**—A slope stability model employing Digital Elevation Model (DEM) data was used in a Geographic Information System (GIS) to predict landslide susceptibility in a large region of the H.J. Andrews Long-Term Ecological Research (LTER) site, located in the Western Cascades of Oregon. To assess the uncertainty of the final output, several different, but equally probable, versions of the input DEM were created through the addition of random, spatially autocorrelated noise (error) files. The DEMs were then processed to produce a family of slope stability maps from which the uncertainty effects of DEM elevation error upon the final susceptibility index could be assessed. The slope stability model itself involved the overlay of rock strengths with slope classes to establish landslide susceptibility indices. The inherent nature of error in DEMs used for modeling slope stability, coupled with much of the Pacific Northwest's commercial timber being located in mountainous terrain, makes this an intriguing problem to solve. The ability to understand the resultant uncertainty due to elevation error when applied to the model has the potential to facilitate improved natural resource management decisions in relation to harvesting and subsequent slope stability.

---

## INTRODUCTION

The US Pacific Northwest contains some of the most productive temperate forests in the world and timber harvesting has been a primary land use in the area for over 100 years. Because most of the commercial timber is in mountainous terrain, such land use activities tend to decrease slope stability and therefore increase the rate of mass movements. One of the major concerns associated with an increase in such movement is the cumulative effects it has on the entire watershed—such as decreased water quality, loss of spawning habitat and organic matter, and debris jams that may break during peak flows, thereby scouring channels and destroying riparian vegetation. Thus, the ability to predict slope failure is a valuable resource in determining the impacts of forestry practices (Swanson *et al.* 1981).

While methods such as ground and aerial surveys can aid in locating existing and potential landslide areas, they also suffer several disadvantages. For example, in the Western Cascade range, ground surveys are difficult to carry out since accessibility is a major problem in a region characterized by steep slopes and very few roads (Swanson *et al.* 1981). This effectively reduces the amount of terrain that can be covered, which in turn may lead to an underestimation of landslide occurrences.

---

<sup>1</sup> Graduate student, Department of Forest Science, Oregon State University, Corvallis, OR

<sup>2</sup> Assistant Professor, Department of Geomatics, The University of Melbourne, Victoria, Australia

Similarly, there can be difficulties in obtaining suitable aerial photography as a result of highly variable weather and climate conditions, seasonal considerations, and problems with photographic interpretation. Unfortunately, landslide inventories in the Western Cascades are best conducted during the winter rainy season and/or immediately after major storms when cloud cover is most prevalent. Also, landslides may be too small to be detected by aerial photography (Swanson and Lienkaemper 1985); or difficult to identify because they are either hidden under the forest canopy (McKean *et al.* 1991), masked in shadow by steep, narrow ridges and tall, dense forest vegetation (Skaugset 1992); or located on streamsid es.

Alternatively, we can use Geographic Information Systems (GIS) to help model landslide susceptibility by applying the critical factors used in determining slope stability. While this approach permits large-scale evaluation of the landscape, it suffers from a lack of knowledge about the uncertainty of the final output resulting from errors occurring in the parameters used to model potential slope failure. The purpose of this paper is to describe how the application of a slope stability model based on DEM-derived values was enhanced to assess the uncertainty in predicting slope failure in the H.J. Andrews Long-Term Ecological Research (LTER) site. The model overlays slope gradient values with a generalized geology map to yield landslide susceptibility classes.

In order to apply the procedure to the slope stability model, the uncertainty assessment technique described in Hunter and Goodchild (1995) was employed in which different, but equally probable, versions of the source DEM were created through the addition of random, spatially autocorrelated elevation error files. The DEMs were then processed to produce a family of slope stability maps from which the effects of error in DEM elevations upon the final landslide susceptibility indices could be assessed. This paper is structured such that a description of the uncertainty model is presented next, followed by discussion of the application of the method to the slope stability model, and finishing with some closing remarks.

## DISCUSSION OF THE UNCERTAINTY MODEL

Before discussing the uncertainty model used in this research, some explanatory remarks are required regarding the term 'uncertainty' itself. In general terms, it denotes a lack of sureness or definite knowledge about an outcome or result and, in the context of GIS, we suggest there is a clear distinction between 'error' and 'uncertainty', since the former implies that some degree of knowledge has been attained about differences between actual results or observations and the truth to which they pertain. On the other hand, 'uncertainty' conveys the fact that it is the lack of such knowledge which is responsible for our hesitancy in accepting those same results or observations without caution, and the term 'error' is often used when it would be more appropriate to use 'uncertainty'.

The uncertainty model that has been used here is a version of the one originally developed by Goodchild *et al.* (1992). In general terms it may be defined as a stochastic process capable of generating a population of distorted versions of the same reality, with each version being a sample from the same population. The traditional Gaussian model (where the mean of the population is an estimate of the true value and the standard deviation is a measure of the variation in observations) is one attempt at describing error, but it says nothing about local variation or the

processes by which it has accumulated. The model applied here is viewed as an advance on the simple Gaussian approach since it has the ability to show spatial variation in uncertainty, as well as the capacity to include in its realizations the probable effects of error propagation resulting from the various algorithms and processes that have been progressively applied to the data sets employed. By studying different versions of the final output, it is possible to see how differences in input affect the outcome, and in essence the purpose of the model could be described as an attempt to “find a Gaussian distribution for maps”.

Elsewhere, in a paper by Hunter and Goodchild (in review), it has been argued that while it is possible to distort or perturb a data set according to an error description (such as the Root Mean Square Error (RMSE) value for a DEM) without any consideration of the likely spatial autocorrelation between point sample elevations, the process may be stochastic but inevitably lacks a certain ‘truthfulness’—since adjacent elevations in a DEM which are otherwise similar in value can be severely distorted, thereby creating large pits and peaks which often do not intuitively occur in nature. This approach produces what are known as ‘random maps’.

On the other hand, the assumption of complete spatial dependence between neighbouring points produces realizations of a DEM which appear ‘truthful’ but not stochastic, since elevations are unnaturally constrained to maintain their relative differences to each other and the introduction of a noise (or error) component has the effect of moving all DEM elevations up or down by a constant amount. Hence, there is a need to find the value in the domain  $0 < \rho < 0.25$  (where  $\rho$  is a measure of spatial autocorrelation) which meets the requirements of being both stochastic and ‘truthful’. The limit of 0.25 ensures stationarity (as discussed in Cliff and Ord, 1981, p. 147) when the Rook’s case is used to test a cell’s elevation against its four neighbours sharing a common edge.

Application of the uncertainty model consists of four stages (as described more fully in Hunter and Goodchild, 1995), with the first stage requiring the user to combine whatever data, processes and models are needed to generate the desired output—in other words, applying the GIS as would normally occur without any consideration of uncertainty. Secondly, the parameters necessary for the realization process are determined by reading system variables associated with the source DEM, such as the number of rows and columns in the data file, the cell size, and geo-referencing details. These will be required later when the elevation noise files are transformed to agree with the original DEM. An error estimate for the DEM will also need to be identified and this is usually a global value of the elevation error present, such as the RMSE as supplied by the DEM producer.

Whilst not a direct step in the realization procedure, the noise files to be employed are usually pre-computed and permanently stored in the system for future use. To date, it has been considered sufficient in the applications tested for about ten files to be held against each  $\rho$  value, with the default values being  $\rho = 0.0, 0.05, 0.10, 0.15, 0.20, 0.21, 0.22, 0.23, 0.24, 0.245, \text{ and } 0.249$ . As for the maximum value of  $\rho$  employed (0.249), experience has shown there is little to be gained from using values higher than this since the realization process becomes so constrained that there is no discernible difference between the realized maps and the original product.

In stage 3 of the methodology, it is expected that users will want to see a small number of trial realizations and the default values of  $\rho$  listed above are applied. A single realization for each value is obtained by first applying the parameters derived from stage 2 to geo-reference and transform the coordinates of the noise grid. Next, the error estimate is applied to map the noise values from a normal distribution of  $N(0,1)$  to  $N(0, RMSE)$ . This adjusted noise file is added to the source data to produce a realization to which the commands employed to create the original GIS product are applied. The realized maps and the differences between the realizations and the original output can be displayed in map or graph form. Finally, in stage 4 of the process the user focuses on a specific  $\rho$  value in the range  $0 < \rho < 0.25$ , and applies a series of noise realizations to produce a set of distorted DEMs for which a family of outputs are created. These enable assessment of the uncertainty in the final product due to elevation error in the original DEM.

## THE CASE STUDY AND DISCUSSION OF RESULTS

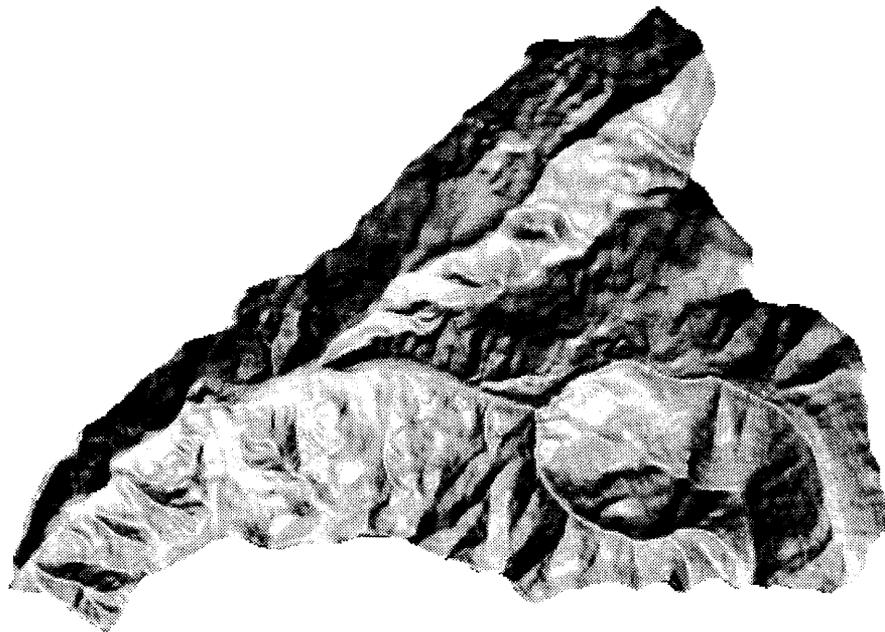
The H.J. Andrews Experimental Forest is located approximately 80 km east of Eugene, Oregon. Lookout Creek Basin (the test area) lies within the forest and drains nearly 6400 hectares in the Western Cascade Range. The overall site was established as an Experimental Forest in 1948 after which trial harvests took place and concomitant road construction occurred throughout the 1960s. Since 1970, site emphasis has shifted to ecosystem research and the levels of timber harvesting and road building have been reduced. The upper areas of Lookout Creek Basin are formed from lava flows ranging in age from 3 to 13 million years, while lower areas are underlain by older (14 to 25 million years BP), more weathered and hydrothermally altered clastic volcanic rocks (Swanson *et. al.* unpublished).

Two types of landslides occur in the basin. The most common are relatively small (averaging approx. 2000 m<sup>3</sup>), shallow, rapid debris slides that take place on hillslopes during intense rainfall and rain-on-snow events. The second type are relatively large, deep-seated, slow-moving landslides (earthflows) that move seasonally at rates varying from centimeters to meters per year. An extensive inventory was conducted which documented over 140 debris slides since 1950 in the basin (Dyrness 1967, Swanson and Dyrness 1975, Swanson *et. al.* 1981). Landslide susceptibility in the basin has been interpreted to be a function of both slope steepness and soil strength, as represented by the type of underlying bedrock (Dyrness 1967, Swanson and Dyrness 1975).

**Table 1.—Derivation of landslide susceptibility classes.**

Slope	Rock Strength	Susceptibility
0° < 10°	Weak	Moderate
	Moderate	Low
	Strong	Low
10° – 20°	Weak	High
	Moderate	Moderate
	Strong	Low
> 20°	Weak	High
	Moderate	High
	Strong	Low

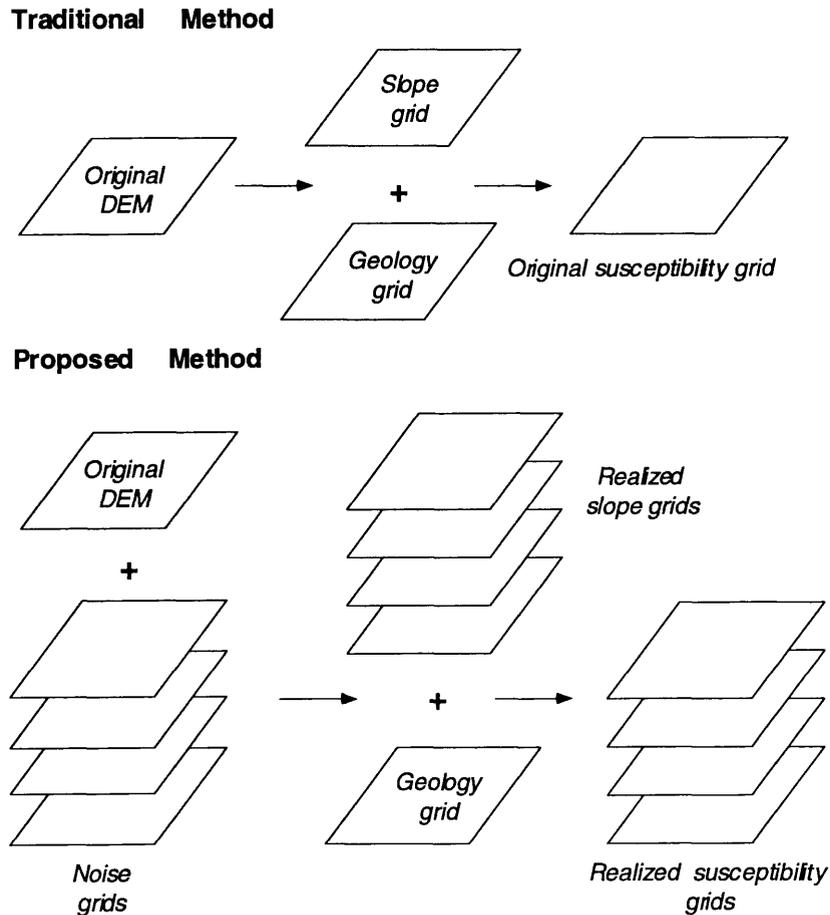
In applying GIS to produce a landslide susceptibility index map, a DEM produced by the U.S. Geological Survey (USGS) was used which has a 30-m resolution and an elevation RMSE of 7 meters. The basin is located within a portion of the DEM measuring 309 x 426 cells and occupies some 69,952 cells or 53% of it (Figure 1). Elevation values in the basin vary from 411 meters above sea level in the south-west to 1615 meters in the south-east.



**Figure 1.—Hill-shaded view of the DEM covering the Lower Creek Basin test area.**

In deriving a landslide susceptibility index map, the slope gradient was first calculated from the DEM for each cell and then reclassified into 3 categories—greater than  $20^\circ$  representing ‘high’ slopes,  $10^\circ$ – $20^\circ$  for ‘moderate’, and  $< 10^\circ$  for ‘low’. A geological map was also created with the same grid cell size and georeferencing as the DEM. Bedrock was classified by strength, with young lava flows and intrusive bodies being graded as ‘strong’; intermediate age clastic rocks including well-welded ashflow units as ‘moderate’; and old hydrothermally altered clastic rocks (18-25 million years BP) as ‘weak’. Using the Arc GRID software throughout the research, the slope gradient and rock strength grids were combined to yield a matrix of nine values, which were then reclassified into three levels of landslide susceptibility—‘high’, ‘moderate’, and ‘low’ (see Table 1). This is the traditional manner in which GIS is used to solve the problem, as shown in Figure 2, and the combination of the DEM with the derived slope grid and geology grid produces a single susceptibility index grid.

Clearly, there is considerable potential here for applying the uncertainty model to assess the cumulative effect upon the final susceptibility index derived for each cell—given that it is a function of the DEM resolution and its estimated elevation error; the slope gradient calculation for each cell; and the method of reclassifying the slope values and rock types to derive the index. In the alternative approach which permits uncertainty to be assessed (Figure 2), elevation noise files with varying levels of spatial autocorrelation were applied to the original DEM to establish corresponding sets of slope files for the test site. The range of ten default values of  $\rho$  as described in the previous section was used. The realized slope files were then taken in turn with the geology file and used to derive ten susceptibility index grids.



**Figure 2.—Comparison of the traditional and proposed methods of calculating landslide susceptibility.**

Each realized grid was subtracted from the original susceptibility grid to give a set of ten difference grids (one for each value of  $\rho$ ), and two graphs were produced showing the mean and standard deviation of the differences respectively for each grid as plotted against  $\rho$ . Both graphs indicated that mean and standard deviation values increased slightly from  $\rho = 0$  to  $\rho = 0.21$ , then dropped suddenly as  $\rho$  approached 0.25. The value of  $\rho = 0.21$  was estimated as the point at which distortion of the DEM produced an output that was both stochastic and 'truthful', however this aspect of the uncertainty model remains subjective and is the subject of continuing research.

It should be noted here that the actual mean and standard deviation values for each realized grid have little significance since they apply to nominal data, even though the susceptibility indices of 'low', 'moderate' and 'high' were assigned numeric values to enable differencing. Nevertheless, the graphs still serve a useful purpose in identifying a  $\rho$  value which can be used for more detailed processing.

The procedure was then repeated by taking nine elevation error grids at the single value of  $\rho = 0.21$  and creating nine realized susceptibility index grids. Since cells exhibiting high instability are of greatest interest, both the original

susceptibility index grid and the realized grids were reclassified so that cells with 'low' and 'moderate' indices were assigned a value of zero, while cells with a 'high' index were allocated a value of unity. The nine realized grids were added together to produce a single grid containing cumulative values denoting the number of times each cell scored a 'high' index on a scale of 0–9. While only nine realizations of the landslide susceptibility index grid were created due to time constraints in preparing this paper, it is considered that a much greater number (such as 100 realizations) would be used in practice to ensure the absence of any bias in the elevation error files used to perturb the DEM. However, the results obtained still enable useful comments to be made about the process.

To visualize the results, we took the original susceptibility map and overlaid all grid cells with a 'high' index (shaded black) on a set of site contours produced at 100 m interval (Figure 3, top). For comparison, we then took the cumulative index grid described previously and selected all cells that scored a 'high' index at least once (Figure 3, middle), together with those that achieved the maximum score of nine (Figure 3, bottom). These cells were also shaded black and overlaid on the same set of contours. Alternatively, we could choose cells on the basis of some probabilistic measure, for example, by selecting all cells which score a 'high' value at least seven out of nine times.

**Table 2.—Relationship between 'high' cell scores and area of the site affected.**

'High Score'	No. Cells	Area of Site
≥1	20,280	29%
≥2	16,843	24%
≥3	14,428	21%
≥4	12,549	18%
≥5	11,034	16%
≥6	9,651	14%
≥7	8,349	12%
≥8	6,872	10%
9	4,982	7%
Original Map	10,168	14%

In the original susceptibility map, the Boolean nature of the slope reclassification into categories with thresholds at 10° and 20° meant that a cell could receive only one index value. However, with the proposed method even though we maintain the same thresholds (and their values are not in question here), we can see how the susceptibility index for a cell behaves under variation of the input DEM. Obviously, cells lying on 'weak' or 'moderate' rock and having slope gradients close to either of the

thresholds are most likely to exhibit variation in their susceptibility index.

Table 2 compares the number of 'high' cells in the original susceptibility grid and the area of test site affected, with the number of cells achieving a 'high' score between 1–9 in the cumulative realized grid. From a user's perspective, selecting cells that achieve a 'high' score at least once is clearly the least conservative choice, given that the elevations of many of these cells will have been perturbed several standard deviations from the mean during the process. We believe this would lead to overestimation of the area affected. On the other hand, selecting only those cells that attain a 'high' score the maximum nine times is the most conservative approach and most likely underestimates the true area of the site susceptible to landslides. There is, however, no right or wrong answer to the selection problem, and the level of risk a user is prepared to accept in the end product becomes a matter of personal choice. In other words, users must learn to live with some degree of uncertainty in their outputs and make value judgements appropriate to individual project requirements.

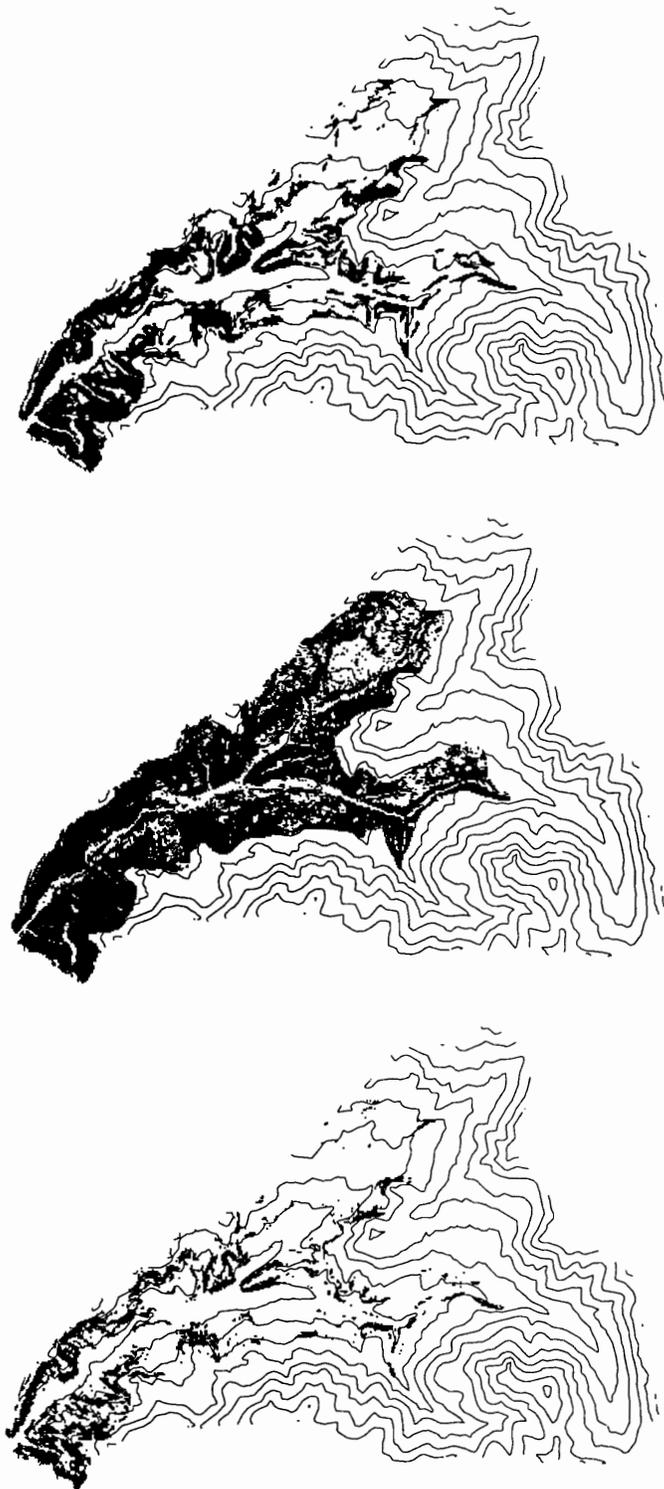


Figure 3.—(Top) All cells in the original susceptibility index map scoring a 'high' value are shown in black overlaid on 100m contours. (Middle) All cells shown scoring at least one 'high' value in the realization process. (Bottom) All cells shown scoring the maximum nine 'high' value.

Finally, there is another error source which will affect the susceptibility index that has not been considered here—positional uncertainty in the geological boundaries used to derive rock strengths. At this stage, the uncertainty model described here is not capable of handling this issue since a solution requires perturbation of the original geological polygons in the horizontal plane. However, research is currently being conducted in this area and another paper elsewhere in this proceedings describes the experimental work that has been done to date to overcome the problem (see Hunter *et al.* 1996).

## CONCLUSIONS

In this paper, the authors have described a process whereby a simple slope stability model used in identifying landslides with the aid of GIS, has been modified to assess the uncertainty in the final susceptibility map as a result of elevation error in the original DEM and its propagation effects upon the intermediate processes involved. Several different, but equally probable, versions of the input DEM were created through the addition of random, spatially autocorrelated noise (error) files, and then processed to produce a family of slope stability maps. Users may then choose the level of risk they are prepared to accept in the final output and identify cells susceptible to landslides that meet their criteria.

There are two benefits to the research. Firstly, the inclusion of uncertainty assessment in the process has the potential to facilitate improved natural resource management decisions in relation to harvesting and subsequent slope stability. And secondly, from a user's perspective the application of the uncertainty model to applications such as the one described here provides a new dimension to their ability to understand the meaning of their GIS products.

## REFERENCES

- Cliff, A.D. & Ord, J.K. 1981. *Spatial Processes: Models and Applications*, Pion, London.
- Dyrness, C.T. 1967. "Mass Soil Movements in the H.J. Andrews Experimental Forest". Research Paper PNW-42, Portland, OR: USDA Forest Service. Pacific Northwest Forest and Range Experiment Station. 12p.
- Goodchild, M.F., Guoqing, S. & Shiren, Y. 1992. "Development and Test of an Error Model for Categorical Data", *International Journal of Geographical Information Systems*, Vol. 6, No. 2, pp. 87-104.
- Hunter, G.J & Goodchild, M.F. 1995. "A Methodology for Reporting Uncertainty in Spatial Database Products". *Urban and Regional Information Systems Journal*, vol. 7, no. , pp. 11-21.
- Hunter, G.J., Höck, B. & Robey, M. 1996. "Experimental Development of a Model of Vector Data Uncertainty". *Proceedings of the 2nd International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Modeling*, Fort Collins, Colorado, 8 pp.
- Hunter, G.J. & Goodchild, M.F. (in review). "Modeling the Uncertainty of Slope Gradient and Aspect Estimates in Spatial Databases". Submitted to *Geographical Analysis*, 22 pp.
- McKean, J., Buechel, S. & Gaydos, L. 1991. "Remote Sensing and Landslide Hazard Assessment". *Photogrammetric Engineering & Remote Sensing*, nol. 57, no. 9, pp. 1185-1193.

- Skaugset, A. 1992. "Slope Stability and General Hydrology Research". *Forest Soils and Riparian Zone Management: The Contributions of Dr. Henry A. Froehlich to Forestry*, pp 23-45.
- Swanson, F.J., & Dyrness, C.T. 1975. "Impact of Clearcutting and Road Construction on Soil Erosion by Landslides in the Western Cascade Range, Oregon". *Geology*, vol 3, pp. 393-396.
- Swanson F.J. & Lienkaemper, G.W. 1985. "Geologic Zoning of Slope Movements in Western Oregon, USA". *Proceedings IVth International Conference and Field Workshop on Landslides*, pp. 41-45.
- Swanson, F.J., Lienkaemper, G. & Nakamura, F. (unpublished). "Change in Landslide Hazard through Time - A Case Study of Forestry Land Use and Climate Change in the Cascade Mountains, Oregon, USA". 15pp.
- Swanson, F.J., Swanson, M.M. & Woods, C. 1981. "Analysis of Debris-Avalanche Erosion in Steep Forest Lands: An Example from Mapleton, Oregon, USA". *Proceedings of a Symposium on Erosion and Sediment Transport in Pacific Rim Steeplands*, Davies, T.R.H., Pearce, A.J. eds., Christchurch, NZ, IAHS-AISH Publication No. 132, pp. 67-75.

### **BIOGRAPHICAL SKETCHES**

Michelle Murillo is a graduate student with the Department of Forest Science at Oregon State University, Corvallis. She gained her BS from the University of New Mexico in Biology.

Dr Gary Hunter is an Assistant Professor with the Department of Geomatics and Acting-Director of the Center for Geographic Information Systems and Modeling, at the University of Melbourne, Australia. He gained his PhD from that university on the subject of handling uncertainty in spatial databases and is the winner, with co-author Professor Michael Goodchild, of the Gerald McCalden Prize at the AURISA 94 conference and the Horwood Critique Prize at the URISA '93 and '95 conferences, for papers dealing with this subject. He is President of the Australasian Urban and Regional Information Systems Association (AURISA) in 1996.

# An Accuracy Assessment of a Spatial Bioclimatic Model

D.W. McKenney<sup>1</sup>, B.G. Mackey<sup>2</sup>, M.F. Hutchinson<sup>3</sup>, R.A. Sims<sup>4</sup>

---

**Abstract.**—Spatially explicit models of forest ecosystems require the integration of biophysical data at a range of scales. Prediction of biotic response should be based on spatially explicit models of the driving environmental processes, in particular those related to climate and topography. Ongoing research in the forests of Ontario have focused on the development of suitably scaled digital elevation models and meso-scaled models of climatic averages. These in turn have been used to spatially extend various biological survey data. In this paper we present some accuracy assessments of the spatial climate models and subsequent differences in species' predicted distributions.

---

## INTRODUCTION

Resource managers require reliable spatial information about species' distributions to help make decisions such as establishing monitoring sites or protected area networks. How important a particular landscape is to a species often depends on where else it does or could occur. Spatial information can help provide context. However considerable effort is required to develop these data. Typical resource inventories such as aerial photographs or remotely sensed data provide relatively little information about non-commercial and/or non-overstorey species, so some other type of modelling is required.

Theory would suggest that ecological response, such as a species' distribution, is a function of a complex myriad of factors including disturbance regimes, human inputs (eg. land use history), biological regimes (eg. interspecies competition) and a number of physical determinants. The latter include thermal, radiation, moisture and nutrient regimes. Data on climate, terrain and soil parent material are the basic inputs to modelling these regimes across landscapes. If spatial data on the

---

<sup>1</sup> Senior Economist, Canadian Forest Service - Sault Ste. Marie, Ontario, Canada.

<sup>2</sup> Senior Lecturer, Department of Geography, The Australian National University, Canberra, Australia.

<sup>3</sup> Senior Fellow, Centre for Resource and Environmental Studies, The Australian National University, Canberra, Australia.

<sup>4</sup> Research Scientist, Canadian Forest Service - Sault Ste. Marie, Ontario, Canada.

regimes are available and species' responses can be calibrated with the regimes, then there becomes an enhanced capability to extrapolate information gained from analysis of field based observations. This type of empirical framework is directly relevant to the problem of spatially quantifying species' distributions, abundance, and productivity (eg. Mackey 1994). In this paper, some accuracy issues associated with a spatial climate model and the problem of extrapolating site data to quantify a species' (jack pine, *Pinus banksiana*) potential distribution are investigated.

## A SPATIAL CLIMATE MODEL FOR THE PROVINCE OF ONTARIO

It is well accepted that climate exerts a strong control on both the distribution and productivity of plant species. However, correlations between various climate parameters and the ecological response of interest is often poor. One reason for this is that weather stations are located well away from field survey points and any estimates of climatic parameters involve some interpolation. Some relatively new methods for interpolating climate provide opportunities for generating more reliable spatial estimates of climatic parameters *ex situ* than previously possible (see Hutchinson 1995, 1987). Applications of these methods to estimate long term monthly mean climate values at points of survey and subsequently explain spatial variation in ecological response have been successful (eg. Nix 1986, Mackey 1994, Yee and Mitchell 1991). The methods make use of partial thin plate smoothing splines (Eq. 1) to interpolate historical weather station data:

$$z_i = f(x_i) + \sum_j^p \beta_j \psi_j(x_i) + \epsilon_i \quad (i=1, \dots, n; j=1, \dots, p) \quad (1)$$

where  $f$  is an unknown smooth to be estimated,  $\psi_j$  are a set of  $p$  known functions and the  $\beta_j$  are a set of unknown parameters which have to be estimated. The  $x_i$  generally represent coordinates in two or three dimensional space. The  $\epsilon_i$  are zero mean random errors. Hutchinson (1995) describes the estimation procedures in detail. In applications to date, climate surfaces have been developed as a function of latitude, longitude and elevation (eg. Nix 1986, Mitchell 1991, Mackey et al. 1996) although any number of independent variables are possible in principal (eg. distance from a major water body). These three parameters are known to be important determinants of long-term climate. The resultant climate surfaces can also be mapped relatively easily in a GIS using software which couples the functions to a Digital Elevation Model (DEM) (see Hutchinson 1993). A DEM is a regular grid of latitude, longitude and elevation.

The main advantage of the thin plate spline techniques over competing methods is that they do not require any *a priori* estimation of the spatial auto-covariance structure. In addition, data smoothing can be optimized by minimizing the generalized cross validation statistic (Hutchinson 1995, Hutchinson and Gessler

1994).

Mackey et al. (1996 in press) used these methods to generate climate surfaces for the province of Ontario as a function of x, y (latitude and longitude) and z (km above sea level). Table 1 summarizes the standard errors of the basic surfaces of monthly mean maximum temperature, monthly mean minimum temperature and monthly precipitation. Besides mapping the surfaces and secondary indices in a GIS by coupling the functions to a DEM, estimates of climatic values can be generated at field survey locations. The latitude, longitude and elevation (x, y and z) of each survey location are required. Hence, estimates of various climate variables can be estimated and climate/plant or plant/animal relations can be subsequently examined. Spatial predictions can be made by linking the statistical functions to spatial databases of the independent variables.

**Table 1.—Number of data points and approximate standard errors of fitted climate surfaces for Ontario.**

<b>Climate Variable</b>	<b>Number of Data Points</b>	<b>Standard Error</b>
monthly mean max. temp.	471	0.13 to 0.26°C
monthly mean min.temp.	471	0.27 to 0.42°C
monthly precipitation (old surface)	471	3.2 to 7.6%
monthly precipitation (new surface)	470 (less Ignace)	3.3 to 8.2%

### **Climate Estimate Errors Associated With Incorrect Field Survey Locations**

Considerable effort has been placed on the collation of existing biological survey data in Ontario so that these methods can be applied. For example, there are now over 4100 forest ecosystem classification (FEC) plots in existence throughout the province (Sims and Uhlig 1992). These data are essentially vegetation surveys in which numerous characteristics were measured in 10 by 10 metre plots and included understorey and overstorey plants, soil properties and some mensurational attributes. Some of the newer plots used GPS technology to record location data. If location information was not already recorded, the original survey records and topographic maps were used to record x, y and z (it is also possible to use a DEM to derive estimates of elevation if only x and y are available).

The importance of accurate location information will vary spatially. For example; steeper climatic gradients will exist where elevation changes are more abrupt, or near very large water bodies such as the Great Lakes. To gauge the importance of location information across the province, an analysis of the location data associated with observations of jack pine was performed. Of the 4100 FEC

plots, 951 contained jack pine. The x,y,z coordinates of each of these were perturbed in two ways. Figure 1 reports an example where the elevation (z) values were increased by 50 metres (elevations above sea level in Ontario range from 0 metres near Hudson Bay to over 610 metres northwest of Lake Superior). New climate values were generated with these increased elevation values and compared to the original estimates. Results for three climatic values (growing season length, maximum temperature of the hottest month and precipitation of the coldest quarter) are shown in figure 1. The original range of values at the 951 plots is also reported in the figure.

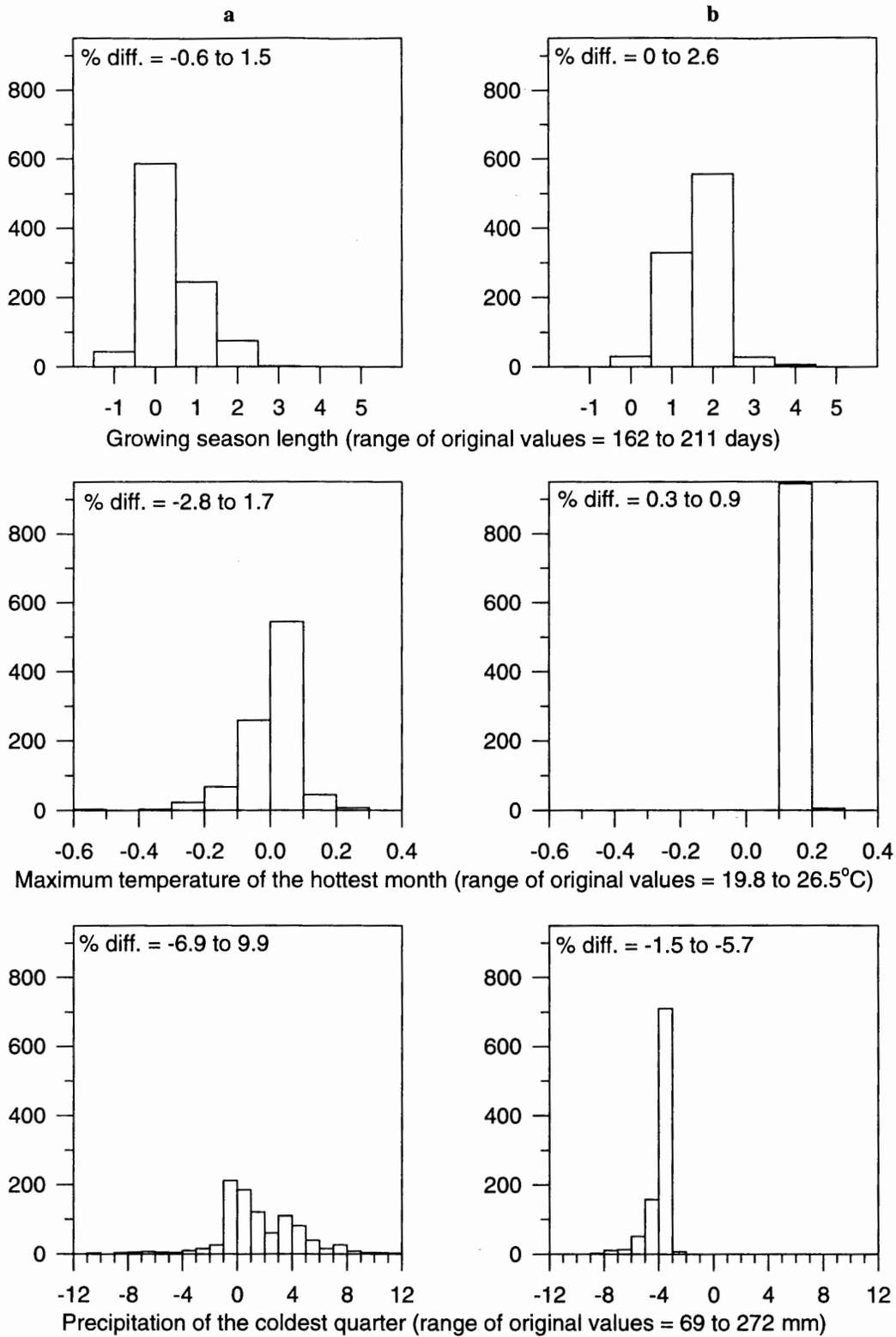
In addition to the elevation perturbations, x and y coordinates were systematically altered. Figure 1 reports the results of increasing x and y by six minutes. Across the range of the plot locations, the implication in terms of distance is approximately 13 km northwest of each "true" geographic plot position.

The changes in elevation and latitude/longitude produced similar changes in growing season length. The changes in temperature and precipitation were consistent with the relative scaling of the input data. The differences from the original values are marginal. The biggest impact was on precipitation in the coldest quarter. In this case the elevation change caused a general decrease in values across all sites (from -1.5% to -5.7% in relation to the original values). The lat/long shift resulted in a -6.9% to 9.9% change in relation to the original values. These differences are for the most part within the standard errors reported in table 1. The results illustrate the spatially varying dependency of climate on these 3 variables.

## **SPATIAL PREDICTIONS OF THE RANGE OF JACK PINE IN ONTARIO**

As mentioned, the ability to generate grids of climatic values and estimates of climate at points of survey offer opportunities to develop spatial predictions of animal-plant/climate relations. A variety of approaches are possible depending on the nature of the biological site data including the use of statistical relationships (eg. presence of an organism =  $f(\text{minimum temperature of the coldest month})$ ). An approach that has proven useful when only presence data is available is a grid matching procedure (Nix, 1986). In this case a climatic profile of the species is generated, based on observations of where it occurs and the estimates of the climatic values at each of those locations. This profile is then matched to grids of those climatic parameters. This is the BIOCLIM/BIOMAP procedure developed by Nix (1986) and colleagues at the Australian National University (McMahon et al. 1995).

BIOCLIM makes use of climate surfaces, DEMs and species' location information to produce bioclimatic profiles. These profiles essentially describe the climatic conditions (eg. the range of mean annual temperatures, annual precipitation) sampled by a plot network. BIOMAP takes the BIOCLIM output



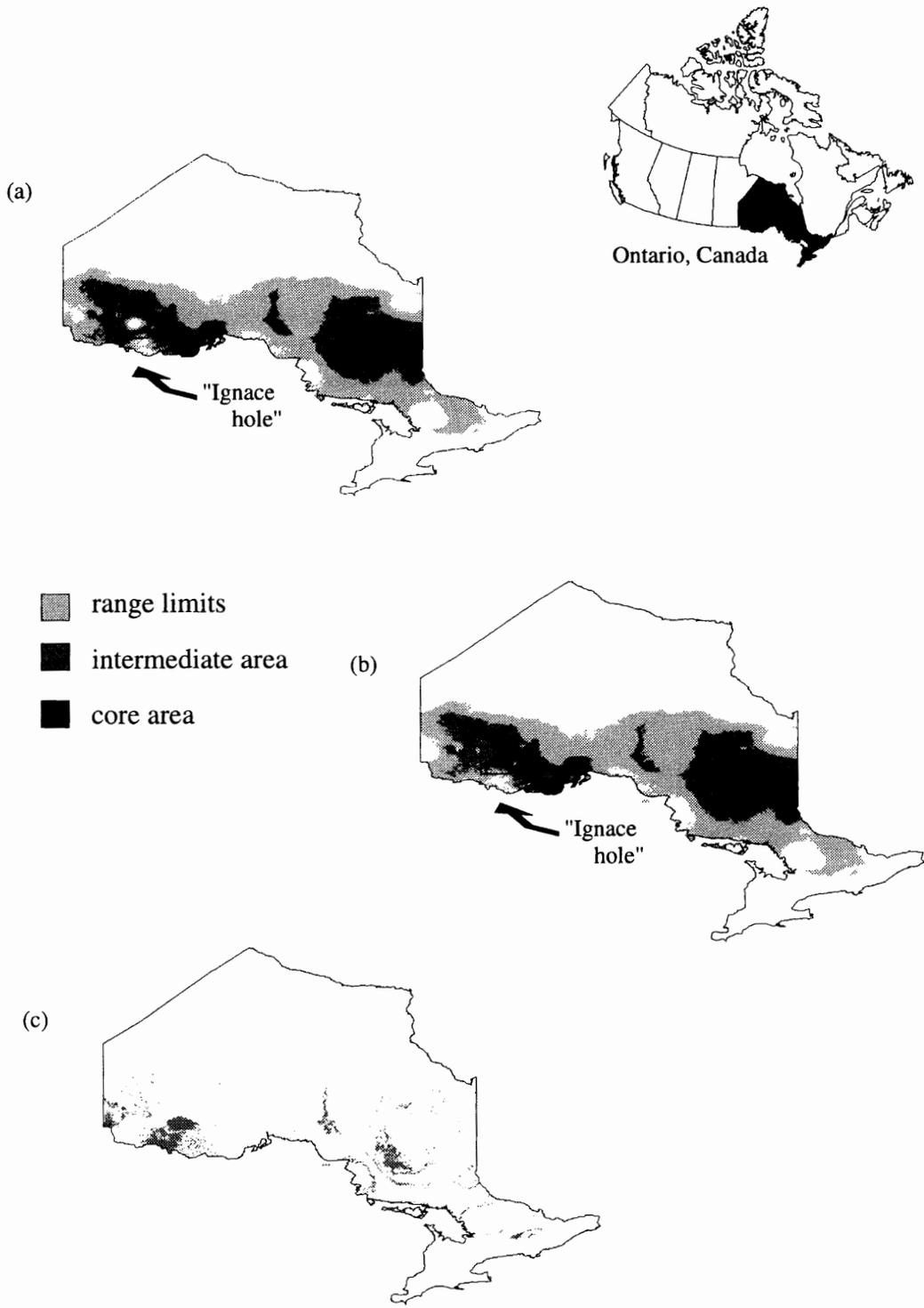
**Figure 1. Differences in three climate variable estimates at 951 FEC plots when perturbing latitude/longitude +6 minutes (column a) and elevation +50 metres (column b). The range of differences as a percentage of original values is presented for each case (% diff).**

and identifies (flags) grid cells which match the bioclimate profile for the selected climatic variables. The resulting data can be input into a GIS and interpreted as a spatial prediction of the climatic domain of the species. In this way, landscapes can be identified which are climatically suitable for the species. Whether the species actually occurs will depend on the influence of other environmental processes.

This approach to predicting the potential domain of a species is relatively transparent to interpretation. However there are limitations; all of the selected climatic variables used in the BIOMAP predictions are assumed to have an equal impact on the species' distribution. Statistical options are limited when the available data indicate only where a species is present; ie. there are no observations of where it is absent. BIOCLIM/BIOMAP is a repeatable and useful method for analyzing these types of survey data and can be easily updated with new observations. However the spatial prediction is also contingent on the representativeness of the underlying biological site data and quality of the climate surfaces.

Figure 2a is a BIOMAP of jack pine based on analysis of 951 province-wide FEC plots. Future work by the authors will provide a detailed discussion of this type of analysis for a suite of commercial tree species in Ontario). A set of bioclimate parameters (eg. annual mean temperature, precipitation in the warmest and coldest quarters) derived from the monthly mean climate variables were coupled to a 1-kilometre DEM of the province in order to produce the BIOMAP. The resulting dark areas represent the "core" climate domain for the species. Grid cells are identified as a core area if all of their climatic parameter values fall within 10-90% of the survey data climatic values (ie. the 10-90 percentiles of the survey climatic parameters are the climatic conditions where most of the site data occur). The lightest shade of grey indicates those grid cells that fall within the range of climatic values defined by the survey data, indicating more marginal conditions.

Noted in figure 2a is a "hole" near Ignace, Ontario, which suggests the area is outside the climatic domain of jack pine. This outcome has been noticed with BIOMAP analysis of several plants and animals and has been a source of some interest given an *a priori* expectation that the region would be quite suitable to many of the species. During an investigation of the spatial pattern of the bioclimatic parameters in that region, it was found that precipitation during the coldest quarter is significantly drier there than the surrounding area. Figure 3a shows the precipitation surface for that area in close detail and also identifies the locations of FEC plots in that part of the province. Clearly there are no FEC plots in this area. Thus, being an empirical approach, the BIOMAP output is doing what would be expected; producing a "hole" at those grid cells where at least one of the climate parameters has values outside the range of the climatic values for known occurrences of the species. An implication is that the area is outside the range of jack pine.



**Figure 2. Climatic domain analysis for jack pine in Ontario using original precipitation surface (a) and new precipitation surface (b) . Map (c) is the difference between (a) and (b).**

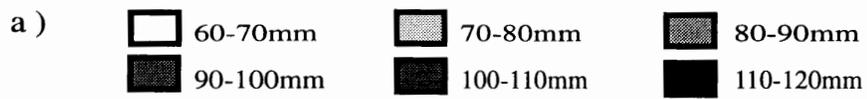
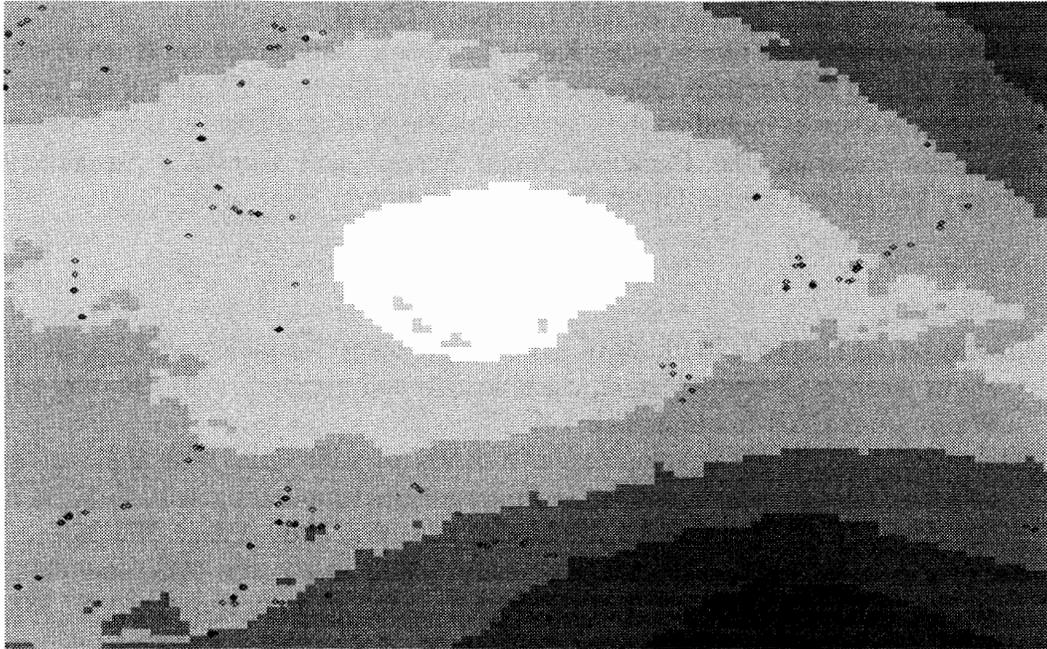
One question is whether the underlying climate surface is robust (see figure 3a). However, the apparent anomaly is a reflection of the input data. It was determined that one weather station (Ignace), at the centre of the "hole" was reporting much drier conditions than nearby stations. For example, long-term November to April precipitation is 138mm compared to 171-212mm for nearby stations. As well as having a shorter history of record collection than other stations, further investigations revealed that Canadian Atmospheric Environment Service experts had several reasons to be suspicious of the quality of data from this particular station. In fact, this Ignace station is now closed.

Given this, a new precipitation surface was created without the Ignace station data. Results are shown in figure 3b. The steep gradient associated with precipitation of the coldest quarter has been smoothed. The standard error of the monthly precipitation surfaces using data from 470 weather stations ranged from 3.3% to 8.2% compared to 3.2% to 7.6% (see table 1). The implication in terms of the predicted climatic range of jack pine is shown in figure 2b. The "Ignace hole" shown in figure 3a is now part of the core climatic range for jack pine. The difference between figures 2a and 2b is shown in figure 2c. Although there are some changes across the province, the most significant changes occurred in the Ignace area where grid cells that were flagged as outside the species range are now part of its core range.

### **CONCLUDING COMMENTS**

Reliable, quantitative spatial data should ideally be used to help make assessments about the potential of landscapes to support particular species. Species distributions must be modelled through integration of various data types and sensitivity analysis of the results to input values should occur. Empirical models are generally driven by the input data, in this paper these are the climate and biological data. The results presented here reiterate one truism well known to ecological researchers: data quality control should be a paramount concern. However it is clear that some data quality issues will not arise without extensive investigation. Prior to the analysis of site data in conjunction with climate, we were quite satisfied with the Ontario climate model. In this case the quest to derive spatial information from site-based observations revealed an area of Ontario that lacked FEC plots and had questionable weather station data.

From a narrower perspective, the analysis summarized in figure 1 suggests considerable scope to make use of historical field data in Ontario. There are numerous vegetation and wildlife survey datasets available in Ontario to which latitude, longitude and elevation could be appended. The analysis of differences in climate estimates in this paper should allay concerns about moderate errors in location records.



b )

**Figure 3. Precipitation in coldest quarter at the 'Ignace hole' (48:45N--50:00N, 91:15W --92:45W) with nearby FEC plots overlaid.**

## ACKNOWLEDGEMENTS

The authors would like to thank Kathy Campbell, Janice McKee, Brian Zavitz and Yin-Quin Yang for their assistance in preparing the figures. Funding has come from various sources including the Canada-Ontario Northern Ontario Development Agreement Program, the Canadian Forest Service and Ontario's Sustainable Forestry Initiative, Ontario Forest Research Institute.

## REFERENCES

- Hutchinson, M.F., 1987, Methods of generating weather variables: *in* Agricultural environments: Characterisation, classification and mapping, A.H. Bunting ed., Wallingford CAB International: 149-157.
- Hutchinson, M.F., 1993, Development of a continent-wide DEM with applications to terrain and climate analysis: *in* Environmental modelling with GIS, M.F. Goodchild, B.O. Parks, and L.T. Stewart eds., Ox. Univ. Press: 392-399.
- Hutchinson, M.F., 1995, Interpolating mean rainfall using thin plate smoothing splines, *International Journal of Geographical Information Systems*, (9): 385-403.
- Hutchinson M.F., and Gessler, P.E., 1994, Splines - more than just a smooth interpolator, *Geoderma*, (62): 45-67.
- Mackey, B.G., 1994, Predicting the potential distribution of rainforest structural characteristics, *Journal of Vegetation Science*, (5):3-54.
- Mackey, B.G., McKenney, D.W., Widdifield, C., Sims, R.A., Lawrence, K., Szczyrek, N., 1994, A new digital elevation model of Ontario, Nat. Resour. Canada - Canadian Forest Service - Sault Ste. Marie ON, NODA/NFP Technical Report TR-6: 26 p.
- Mackey, B.G., McKenney, D.W., Yang, Y., McMahon, J.P., Hutchinson, M.F., 1996, Site regions revisited: A climatic analysis of Hills' site regions for the province of Ontario using a parametric method, *Canadian Journal of Forest Research*, (26): 333-354.
- McMahon, J.P., Hutchinson, M.F., Nix, H.A. and Ord, K.D. 1995. ANUCLIM User's Guide. Draft Report, Centre for Resource and Environmental Studies, Australian National University, Canberra.
- Mitchell, N.D., 1991, The derivation of climate surfaces for New Zealand, and their application to the bioclimatic analysis of the distribution of Kauri (*Agathis australis*), *New Zealand Journal of Royal Society*, (21): 13-24.
- Nix, H.A., 1986, A biogeographic analysis of Australian elapid snakes: *in* Atlas of elapid snakes of Australia, R. Longmore ed., Australian flora and fauna series No. 7. Australian Government Publ. Service, Canberra Australia: 4-15.
- Sims, R.A., Uhlig, P., 1992, The current status of forest site classifications in Ontario, *Forestry Chronicle*, (68):64-77.
- Yee, T.W. and Mitchell, N.D., 1991, Generalized additive models in plant ecology, *Journal of Vegetation Science*, (2): 587-602.

# An Assessment of the Horizontal Accuracy of Interim Terrain Data

L.A. Fatale<sup>1</sup>, J.A. Messmore<sup>2</sup>, and J.R. Ackeret<sup>3</sup>

---

**Abstract.**--Within Army and throughout the private sector, the accuracy of Digital Topographic Data (DTD) is often taken for granted. Users assume terrain data to be accurate merely because of its digital nature. This paper describes a study to determine the horizontal accuracy of Interim Terrain Data (ITD), a 1:50,000 scale terrain analysis data base. Due to its interim nature, ITD may be digitized from hard copy source ("carto-controlled") or produced solely from photography ("photo-controlled"). The producer (Defense Mapping Agency, DMA) provides no definitive accuracy statement with the data. Over 400 ITD features were chosen for analysis. Each feature was located in the field using a Precise Lightweight Global Positioning System Receiver (PLGR) in the Precise Positioning Service (PPS) mode ( $\approx 10$  meter accuracy). Universal Transverse Mercator (UTM) coordinates obtained from the PLGR were annotated for each feature. UTM coordinates for coincident features in the ITD digital files were subsequently extracted by GIS query in order to statistically compare the digital data to the field data, calculate the offsets, and thereby determine the horizontal accuracy of the ITD. Potential accuracy differences between the carto-controlled and photo-controlled ITD and the variability of natural vs. man-made features are also discussed.

---

## INTRODUCTION & BACKGROUND

Interim Terrain Data (ITD) was first produced by the Defense Mapping Agency (DMA) in 1987. Production of this data was initiated to fill the Army's urgent need for a digital terrain analysis product in support of emerging topographic and Command and Control (C2) systems. Earlier, in 1984, the Army had documented a digital topographic data (DTD) requirement that encompassed all Army's known and anticipated tactical DTD requirements. However, production against this requirement, described as Tactical Terrain Data (TTD), was not scheduled to commence until 1996. In order to get similar data to Army users, DMA initiated

---

<sup>1</sup> *Physical Scientist, U.S. Army Topographic Engineering Center, Alexandria, VA*

<sup>2</sup> *Supervisory Physical Scientist, U.S. Army Topographic Engineering Center, Alexandria, VA*

<sup>3</sup> *Physical Scientist, U.S. Army Topographic Engineering Center, Alexandria, VA*

production of an interim product called ITD, the goal of which was to fill the data void until TTD production came on line. Due to the interim nature of this data set and the use of more than one production method, DMA does not state a formal horizontal accuracy measure for this product.

As the center of expertise for DTD within the Army, the Digital Concepts and Analysis Center (DCAC) at the U.S. Army Topographic Engineering Center (TEC) formulated this study to characterize the horizontal accuracy of ITD. Knowledge of horizontal accuracy is important to the Army as well as other DOD users because it affects what are appropriate uses of the product. For example, is ITD as accurate as the Topographic Line Map (TLM), which may be used for targeting purposes, or is it only as accurate as a city graphic which is not an appropriate source for targeting? Fortunately, along with the need to know ITD's horizontal accuracy, technology for rapid position determination has become available in the form of Global Positioning Systems (GPS). This technology was applied to the determination of the horizontal positions of well-defined ITD features.

The ITD is produced through two production processes, one is a hard-copy digitizing process, and the other is a pure photogrammetric production process. In the hard-copy process, tactical terrain analysis data base (TTADB) mylar overlays (which are geo-registered to the corresponding 1:50,000-scale TLM's) are digitized, thereby resulting in a "carto-controlled" ITD product. The photogrammetric production process is effected using imagery on DMA's feature extraction (FE) workstation, thereby resulting in a "photo-controlled" ITD product. It should be noted that ITD cells from both of these widely divergent production processes are currently distributed and in broad use. Carto-controlled cells make up the bulk of currently available ITD ( $\approx 840$  of 1340 cells). The  $\approx 500$  remaining cells are photo-controlled.

## **PURPOSE**

The purpose of this study was to provide the Mapping Charting & Geodesy (MC&G) user community with information about the horizontal accuracy of ITD features. It is intended to give the user a level of confidence in regard to ITD accuracy and to establish the utility of ITD for various applications. The study also identified accuracy differences between cartographic and photogrammetrically controlled ITD.

## **METHODOLOGY & FIELD COLLECTION PROCEDURE**

To facilitate DCAC's ITD accuracy investigations, ITD hard-copy plots and digital data for 10 cells in the continental U.S. (CONUS) were obtained from DMA along with the associated TLM's. The use of the hard-copy plots (from digital ITD) instead of the TTADB ensured that features identified for analysis would be represented in the digital data. Each of these plots covered approximately a 15-by

15-minute area. There were four photo-controlled cells in the Alamogordo/Las Cruces, NM area as well as six carto-controlled cells; two in the Ft. Bragg, NC, area and four in the vicinity of Ft. Hood, TX.

Scientists at DCAC reviewed the hard-copy plots and identified appropriate features for field reconnaissance. Two factors weighed heavily in the selection of features. The first factor was the existence of a distinguishable point on the ground (e.g., a road intersection, bridge, dam, or fence line). The second factor was accessibility for measurement of its horizontal placement. No vertical accuracy criteria, such as feature height or height above mean sea level, were considered. The selected features were then annotated on the coinciding 1:50,000 TLM for field use. They comprised a representative sample from the Transportation (TRANS), Surface Drainage (SD), Obstacles (OBS), and Vegetation (VEG) overlays since these contained point or line features that were more easily recognizable in the field. Features in the Surface Configuration and Soils overlays were not utilized for this study. Although important for many applications, these features are not easily distinguishable on the ground and are difficult to measure accurately.

Depending on the density of the data, between 14 and 73 features on each of the 10 cells in the study areas were chosen for analysis. The number and variety of features chosen were those deemed necessary to assure representative coverage throughout each cell. Over 400 features were subsequently visited in the field. Each site was located using a Precise Lightweight Global Positioning System Receiver (PLGR) in the Precise Positioning Service (PPS) mode ( $\approx 10$  meter accuracy). Universal Transverse Mercator (UTM) coordinates derived from the PLGR were annotated for each.

## **ANALYSIS**

### **Data Preparation**

The digital ITD files provided by DMA were partitioned on the provided media (9-track tape/6250bpi) by individual sheet number and thematic layer(s). Related file header data was used to acquire descriptive information (metadata) pertaining to the individual data sets. The digital ITD files were archived in Standard Linear Format (SLF). Arc/Info (V. 7.0) was used to import the ITD-SLF files read from the tape. Upon import, the data was converted into Arc/Info coverages using the "SLFARC" command. Arc/Info's GIS functionalities enabled us to:

1. Identify & display selected features by thematic layer, feature type, and feature attributes.
2. Precisely extract UTM coordinates for selected features.

The Arc/Info coverages were geo-referenced in WGS-84 spherical geographic coordinates (LAT- LONG). Each coverage, maintained in single-coordinate precision (seven significant digits), was subsequently projected into a WGS-84 planar coordinate system (UTM) which was consistent with the GPS field collection coordinate system. Projection of coordinates from LAT-LON to real-world UTM coordinates via the Arc/Info projection utilities was determined to be a virtually error free process. Since both input and output coverages were stored in WGS-84 and required no additional datum transformations, the statistical significance of the root-mean-square (RMS) error was  $\leq 0.010$ . [NOTE: A perfect projection with an RMS error of 0.000 is not possible with real-world data.] Customized graphical user interfaces were developed utilizing Arc/Info's Arc Macro Language (AML) to facilitate extraction of digital ITD feature coordinates.

### **Digital ITD Coordinate Extraction**

UTM coordinates for selected features were extracted in order to compare the field data to the digital data, calculate the offsets, and thereby determine the horizontal accuracy of ITD. Thematic coverages for each cell in the study area were displayed with the features appearing as unsymbolized center-line data. Next, those ground features visited in the field were interactively identified in the digital file. A coordinate pair was extracted and recorded alongside the coordinate pair of the coincident ground feature. Finally, a series of statistical analyses were performed on both sets of coordinates.

### **Statistical Analyses**

Coordinates for the photo-controlled and the carto-controlled ITD features were stored in different files which were further delineated by eastings and northings. These files were then grouped by specific feature types (e.g., "surface drainage features from photo controlled data"). Standard statistical measures such as mean, standard deviation, 95 percent confidence interval and a t-statistic with an associated critical value were calculated for each data file. The mean and standard deviation determine the size of the error. The 95 percent confidence interval indicates the potential magnitude of that error in future measurements. The t-statistic determines whether there is a significant difference between two populations. The critical value is the threshold that the T-statistic must exceed, at a given confidence level and sample size, to ensure rejection of the null hypothesis (i.e., for this study, that there is no difference between the field data coordinates and those of the ITD data).

## DISCUSSION/FINDINGS

During the field work, DCAC scientists visited 421 feature locations, collecting an easting and a northing at each for a total of 842 observations (combined photo- and carto-controlled data). A summary of statistics for this data is found in Table 1. Offsets ranged from a low of 0 meters (perfect correspondence between field & ITD coordinates) to a high of 339 meters. The average offset for all observations was 25 meters. When the GPS measurement error (Q value) of 10 meters is included, an average offset of 15 to 35 meters is established.

**Table 1.—Summary of ITD Feature Coordinate Offsets**

<i># obs. (E&amp;N)</i>	<i>OFFSET (m) Low Mean High</i>	<i>Std. Dev. (m)</i>	<i>#% offset 50m</i>	<i>#% offset 100m</i>	<i>Q</i>
842	0 25 339	31	73/8.67	18/2.14	10

Q - measure of GPS error (circular) in meters

E&N - Eastings & Northings

Of the 842 total observations, 769 features (91.3 %) were within 50 meters of their expected locations. This is within the expected accuracy of the 1:50,000 Class B TLM (50 meters circular @ 90% confidence), which is the only related product with a formal accuracy statement. Only 73 features (8.7%) had field/ITD offsets above the 50-meter TLM baseline. Further examination revealed that 18 (2.1%) of the 73 had offsets over 100 meters. As expected, 93 percent (68 of 73) of the offsets over 50 meters occurred in cartographically controlled data. These errors were especially prevalent in the SD (e.g. dams, canals), OBS, and VEG coverages where 55 of the 73 offsets were observed. Potential rationales for the higher offsets in the carto-controlled SD, OBS, and VEG coverages (vs. photo-controlled ITD) lie in DMA's collection/symbolization techniques for this data as described below.

In carto-controlled ITD, horizontal positioning is keyed to a base map. Although many features are collected from imagery, they are still subject to a potential offset of 50 meters (map accuracy). Moreover, SD features such as dams and canals, vegetation polygons, and especially obstacles are often displaced for symbolization purposes. Obstacles such as embankments, escarpments, ditches or road cuts may be substantially displaced to ensure the correct alignment of a stream or road bed. Moreover, it is often difficult to determine the exact start or end point of an obstacle in the field.

Vegetation features are inherently characterized by their natural variability. Furthermore, ITD vegetation has a minimum areal size requirement resulting in the inclusion of many different vegetation types within a single polygon. The resulting boundaries can vary in these transition areas causing offsets.

The production procedure for delineation and depiction of dams and canal end/break points (above/below ground transition) is unlike that of other SD features in which positional accuracy is strictly upheld. Dam and canal lengths are often variable since they are roughly plotted to scale to provide a relative size. More emphasis is placed on the location/axis of these features rather than their size/length.

Figure 1 depicts the data offset for an actual point bridge as observed on the Gatesville, TX cell, map sheet 6446-1. The ITD bridge feature is approximately 25 meters from the road/stream crossing point where the field measurement took place. However, the actual position of the bridge may be as close as 15 meters or as far as 35 meters from its intended location depending on the GPS error introduced in the field.

Table 2 illustrates various comparisons between the carto-and photo-controlled ITD at the thematic level. Cartographic feature offsets are acceptable, with some coverages performing better than others. For example, offsets for the carto transportation features averaged approximately 24 meters, while obstacles offsets approached 70 meters. The mean offset for the carto vegetation features ( $\approx 23\text{m}$ ) was comparable to transportation, but exhibited a much higher 95 percent confidence interval ( $\approx 22.5$  vs.  $\approx 4.5\text{m}$ ). This indicates increased variability for the vegetation features and a potentially lower correlation with subsequent field observations.

Overall confidence intervals for the carto-controlled features ranged from  $\approx 4\text{m}$  (transportation) to  $\approx 53$  meters (obstacles). The mean offset for the cartographic data is 28.8 meters. Conversely, the mean offsets of the photo-controlled features were very low across all thematic layers. They ranged from a low of 10-12 meters (vegetation, surface drainage, and transportation features) to a high of 21 meters (obstacles). The 95 percent confidence intervals are fairly consistent across features ranging from  $\approx 6$  meters (transportation) to  $\approx 22$  meters (obstacles). The mean offset for the photo data is 13.7 meters. This is directly attributable to the increased fidelity provided by a solely photogrammetric production technique.

It is recognized that statistics/comparisons for several of the feature groupings in Table 2 are based on small sample sizes ( $<30$  observations). However, in each case the t-statistic exceeds the critical value, indicating that random error cannot explain the variance observed between the field data coordinates and those of ITD. Therefore, there is a significant difference between the populations at the 95 percent confidence level.

It should also be noted that all of the ITD data used for this study was collected over areas of the continental United States (CONUS) with relatively reliable source materials. Assuming the availability of comparable sources, findings in this report should be extensible to any location. However, it is unknown to what extent, if any, foreign or other ancillary source materials may affect the results herein.

**Table 2.--Carto/Photo ITD vs. Field Data Offsets by Feature & Associated Statistics**

*Photo-Controlled Data in meters*

<i>Measurement Type</i>	<i>Average (Mean)</i>	<i>Std. Dev.</i>	<i>95% Confidence Int. Of Mean</i>	<i>T Stat.</i>	<i>Crit. Val.</i>	<i>Sample Size</i>
All features	<b>13.72</b>	19.49	10.98 - 16.47	9.8	1.65	196
Transportation features	<b>12.69</b>	17.29	9.69 - 15.69	8.37	1.66	130
Obstacle features	<b>21.31</b>	30.85	10.19 - 32.44	3.91	1.70	32
Surf. Drainage features	<b>11.00</b>	11.57	5.25 - 16.75	4.04	1.74	18
Vegetation features	<b>10.00</b>	9.56	4.91 - 15.09	4.19	1.75	16

*Carto-Controlled Data in meters*

<i>Measurement Type</i>	<i>Average (Mean)</i>	<i>Std. Dev.</i>	<i>95% Confidence Int. Of Mean</i>	<i>T. Stat.</i>	<i>Crit. Val.</i>	<i>Sample Size</i>
All features	<b>28.76</b>	32.25	26.27 - 31.25	22.67	1.65	646
Transportation features	<b>24.43</b>	24.97	22.20 - 26.65	21.61	1.65	488
Obstacle features	<b>69.68</b>	75.57	43.30 - 96.05	5.38	1.69	34
Surf. Drainage features	<b>37.04</b>	30.05	31.14 - 42.94	12.45	1.66	102
Vegetation features	<b>23.27</b>	25.42	12.00 - 34.54	4.29	1.72	22

## CONCLUSIONS

1. The offset of well-defined ITD features taken as a whole was approximately 25 meters. Approximately 90 percent of these features were within 50 meters of their expected locations. This meets the accuracy specification of 50 meters for a class B, 1:50,000-scale TLM, a traditional source for targeting information.
2. Overall, carto-controlled ITD feature offsets (29 meters) were higher than photo-controlled ITD feature offsets (14 meters). Our expectations were that the photo-controlled ITD would perform very well, as it did; however, performance of the carto-controlled ITD was also suprisingly good.
3. The positional accuracy of ITD features was determined to be compatible with the positional accuracy of hand-held precise GPS equipment currently used in the field. This accuracy compatibility will lessen the chance for confusion and correlation problems such as could occur when features are located close to one another on the ground and GPS-derived locations do not match data base locations. This problem may be more of a concern with lower accuracy data sets such as VMAP Level 1 (1:250,000 scale) and DCW (1:1,000,000 scale).

## ACKNOWLEDGMENTS

The authors wish to extend special appreciation to Mr. John Nedza, Geographic Information Lab, Topographic Engineering Center (TEC) for the many hours of effort he provided during import of the raw Interim Terrain Data (ITD) data and subsequent conversion to Arc-Info coverages. His outstanding technical expertise greatly facilitated data analyses and played a critical role in the successful completion of this study.

The authors also wish to thank Mr. Bill Ryder, Special Studies Division, Digital Concepts and Anaysis Center, TEC for his excellent terrain analysis support during the field collection; Professor Richard Scott, Rowan College, Glassboro, N.J. for advice and insight about the statistical processes used in the report; and Messrs. Robert Gormley and David Wu, Special Studies Division, Digital Concepts and Analysis Center, TEC for their work in regard to research and compilation of the statistical analyses.

# Acquisition of Spatial Data by Forest Management Agencies: a Review

Michael J.C. Weir<sup>1</sup>

---

**Abstract.** — Forest management agencies have a long tradition of surveying and mapping for the acquisition of a wide range of spatial data. These include: i) external (cadastral) boundaries; ii) permanent, internal, forest boundaries; iii) forest stand and other impermanent boundaries; and iv) roads and other topographic features. In recent years, the well-established processes of graphical map compilation and revision have been increasingly replaced by digital mapping techniques.

Based on a limited survey, together with published information, this paper reviews the surveying and mapping techniques currently employed by forest mapping agencies for the task of digital database establishment and maintenance. Accuracy tolerances specified for different categories of information vary considerably. The survey showed that not all agencies apply checks to ensure that specified tolerances are achieved. It appears also that some agencies apply techniques whose achievable accuracy is inconsistent with specified tolerances.

---

## INTRODUCTION

The acquisition of spatial data to assess the state of forest resources has its origins in central Europe during the 15th century (Griess, 1989). Since the early 19th century, when forest management based on reliable area statistics gradually became common practice in Europe, forestry agencies have employed personnel and established facilities for the specific tasks of surveying and map production. In Canada alone, about 30 million hectares are inventoried annually by provincial forestry agencies. These agencies maintain more than 37,000 map sheets at scales varying from 1:10,000 to 1:20,000 (Leckie and Gillis 1995). Spatial data acquisition by forest management agencies worldwide therefore accounts for a significant portion of mankind's total effort and investment in resource mapping.

The aim of this paper is to present a review of the surveying and mapping techniques currently employed by forest mapping agencies for digital database establishment and maintenance. The presented information is mainly based on data

---

<sup>1</sup> Lecturer, International Institute for Aerospace Survey and Earth Sciences (ITC), Enschede, The Netherlands.

from a small survey of forestry agencies known to use GIS in production (research organizations, were not included). The aim of this survey was to determine: i) how forest management agencies acquire spatial data for input to GIS; ii) the accuracy tolerances (if any) applied; and iii) policies concerning the updating of the spatial database. A questionnaire covering these topics was sent to 33 agencies in various parts of the world and 25 replies were received.

The remainder of the paper is divided into three sections. The first section briefly describes the main categories of spatial data required by forest management agencies. The technology used by the 25 surveyed agencies to acquire these data is summarized. In the second section, specified accuracy tolerances are examined in the light of the surveying and mapping technology employed. The paper concludes with some remarks on spatial data accuracy and error modelling in forest management.

## **SPATIAL DATA ACQUISITION**

Although maps are probably the documents most frequently consulted during forest operations (Moser 1970), technical details their production are not widely published, neither in the forestry nor in the surveying and mapping literature. Leckie and Gillis, (*op. cit.*) provide extensive details of current forest mapping practice in Canada. A recent report (Kennedy *et al.* 1994) summarizes forest mapping activities in Europe. Pröbsting (1994) provides a useful review of applications of aerial photography in European forest mapping.

### **Information content**

The management of forests and forest land requires a wide range of spatial information, usually obtained from a variety of sources. For example, a GIS to support the formulation of "Indicative Forestry Strategies" in Scotland uses 34 sources of biophysical and administrative data (Aspinall *et al.* (1993). Although requirements vary from agency to agency depending on forest conditions and management goals, maps and GIS databases generally comprise four main categories of spatial data, namely: cadastral boundaries, "internal" administrative (for example, compartment) boundaries, topographic details and information on the forest resources. Information on spatial data acquisition by the 25 management agencies was compiled for each of these categories of information (table 1). It should be noted that the data in the table refer only to the number of instances where a particular technique is mentioned as being used. The data do not consider the area concerned or the amount of equipment employed. Furthermore, many agencies apply more than one technique to acquire certain categories of spatial data.

**Table 1. - Sources of spatial data required by 25 forest management agencies  
(number of occasions mentioned).**

Source	Cadastral boundaries	Topographic features	Internal boundaries	Forest boundaries
Supplied by other agencies as:				
maps	12	15	5	2
digital data	8	6	1	1
Acquired by the forest agency using:				
Land survey	7	6	9	9
GPS	1	5	1	4
Graphical transfer from aerial photographs	1	6	6	13
Graphical stereoplotting	1	2	4	6
Digital stereoplotting	1	5	4	7
Orthophotography	3	7	4	6
Satellite imagery	-	2	1	2
Not required	1	-	6	-

### **Cadastral information**

An exact knowledge of the area and limits of the forest land is essential for proper management. In many parts of Europe, large scale cadastral maps are the source documents for the compilation of the "basic forest map" (German *Forstgrundkarte*). Forest management agencies generally obtain information on land ownership boundaries from cadastral survey agencies. In some countries, however, the forest management agency itself may have official cadastral survey tasks such as the determination and demarcation of parcel boundaries on forest land (Wander 1981).

### **Topographic information**

General topographic information is required for planning forest operations such as road construction and timber harvesting and also to provide a metric framework within which sample plots for forest inventory can be located. Most importantly, details such as roads or drainage features may define (parts of) forest stand boundaries. Topographic information therefore forms an integral part of any (graphical or digital) forestry database rather than simply a background against which details, such as forest types, are mapped. The survey of 25 forest agencies showed that they obtain topographic information primarily from other agencies. However, surveys by forestry agencies themselves are generally required to map additional details not shown on general purpose topographic maps and "new" features such as recently constructed forest roads. Land surveying, GPS and a

range of photogrammetric methods are all employed for this purpose. When relatively approximate methods of graphical transfer from aerial photographs are employed, these additional topographic features are likely to be acquired with a spatial accuracy lower than that achieved by the national mapping agency, leading to inconsistencies in spatial data quality among different sources.

### **Permanent boundaries established by forest agencies**

Many forestry organizations establish and maintain a "quasi cadastral" survey of the land under their jurisdiction. This may take the form of compartments or timber concession areas, the boundaries of which must be demarcated, surveyed and mapped. The survey showed that a wide range of surveying and mapping techniques are employed to acquire data on permanent "internal" forest boundaries. Unlike "external" property boundaries which, for reasons of accuracy, are primarily surveyed in the field, permanent forest boundaries need not be surveyed to cadastral mapping standards. Photogrammetric techniques are therefore feasible and, in fact, apparently predominate over field surveys.

### **Forest information**

Information on stand boundaries and other "specialized" forest information is not generally available from existing sources. Acquisition of this category is therefore one of the main tasks of any forest surveying and mapping agency. Just over half of the agencies employ simple methods of graphical transfer from aerial photographs for this purpose. The operating costs of these "simple" graphical methods are not significantly lower than those associated with the use of rigorous photogrammetric methods (Weir 1981). It is therefore surprising to find this labour intensive approach still so widely used by forestry agencies in developed countries. This technique is employed by many provincial forestry agencies in Canada (Leckie and Gillis, *op. cit.*), although some are now replacing it by more efficient photogrammetric methods such as digital mono plotting.

As table 1 indicates, forest management agencies make only limited use of satellite imagery for forest type mapping. Clearly, the spatial resolution of the present generation of earth observation satellites is inadequate for mapping at the scale, level of detail and positional accuracy generally required by agencies practicing intensive forest management. In many developing countries, however, satellite imagery may be the only source of timely spatial data on forest resources, although informal discussion with foresters from these countries suggest that there too, aerial photography remains the principal source of spatial data.

### **Frequency of revision**

Unlike a soil or geological map which, if carefully produced, may serve as a useful document for many decades, a forest management map must be updated at

regular intervals. Consequently, map revision is an important aspect of forest management mapping. In the intensively managed forest of western Europe, where management plans are traditionally prepared every ten years, management maps are revised on a corresponding 10-year cycle.

Nowadays, map revision is increasingly carried out as a process of (digital) spatial database maintenance. Table 2 gives details of the frequency of updating employed by 16 of the surveyed agencies and indicates a tendency towards continuous or annual updating. Corrie *et al.* (1994) describe the database maintenance procedures employed by a timber company for 2.4 million hectares of productive forest in British Columbia. The database, which includes 250 base and thematic layers, is continuously updated, with all changes distributed to local offices on a nightly basis.

**Table 2. - Frequency of database updating by 16 forest management agencies.**

	<b>Cadastral boundaries</b>	<b>Topographic features</b>	<b>Internal boundaries</b>	<b>Forest boundaries</b>
Continuous	7	9	9	5
Several times per year	-	1	-	1
Annually	3	3	3	4
Every 2 - 5 years	2	1	1	1
Every 10 years	2	1	3	5

In fact, experience with database updating is still limited. At the time of the survey (mid-1994), only two systems had been in operation for more than ten years. Eight agencies had been using their system for less than five years and some agencies had not yet done any updating. Interestingly, a few agencies indicated that forest stand information is updated by means of a complete remapping rather than by considering only the actual changes. Although this would appear to negate the advantages of having a digital database, there may be a limit to the amount of revision which can take place. One agency, which has been employing GIS since 1977, updates the database on a 2-year cycle and makes completely new overhaul every ten years.

### **ACCURACY STANDARDS**

In forest management, spatial data are required to provide not only locational information but also, most importantly, area data for the calculation of timber volume and other quantities. Ideally, therefore, the errors associated with these area data should be known. Leckie and Gillis (*op. cit.* p81) report that most Canadian forest inventory maps are considered to have accuracies "in the 10 to 25m range", although as Gemmell *et al.* (1991, p221) note, "most existing forest

inventories do not have estimates of the accuracies of their polygon labels and boundaries”.

The survey of 25 forest management agencies suggested that, among the four main categories of information, cadastral boundaries tend to have relatively stringent tolerances. As indicated in table 1, however, this category of data is generally provided by other agencies according to predetermined standards. The other three categories of information are mostly all acquired to the same standards (although some agencies relax the tolerances for - relatively uncertain - forest stand boundaries). Actual values vary markedly from one agency to another or are expressed as a range (for example, 10 to 30m) within which map accuracy should fall. Tolerances for stand boundaries set by individual forest management agencies in Europe are generally in the range 2 - 10m rmse relative to nearby fixed detail.

Although virtually all of the 25 surveyed agencies indicated that they set accuracy standards for the acquisition of spatial data, almost half (12) indicated that they do not apply any quality control measures to check if accuracy tolerances are actually being met. Interestingly, eight of these 12 agencies employ simple graphical methods for mapping forest boundaries from aerial photographs. Tests using a variety of instruments, photo-scales and operators (Weir, *op. cit.*) have shown that the accuracy of these methods is highly sensitive to terrain conditions, base map quality and operator skill. Their performance (under optimal conditions, 0.8 - 1.3mm rms plotting error) is generally inconsistent with required standards.

## CONCLUSIONS

This paper has shown that many forestry agencies already work to, or at least aim at, “near professional” standards of surveying and mapping. In some cases, however, there is a need to define these standards more clearly and, in particular, to apply appropriate quality control measures.

In forestry, spatial data acquisition is not simply a matter of map production. Data on the areas of individual forest stands form an essential input for forest management planning. Spatial data quality therefore has a direct impact on forest management operations and business decisions (Prisley 1994). The establishment and maintenance of a spatial database can represent a considerable investment for a forestry agency, and it may be difficult to justify measures to improve spatial data quality (Keefer 1994).

In recent years, GIS researchers have laid much of the ground work needed to handle spatial error in natural resource databases. A lot of the research on error propagation (see, for example, Heuvelink 1993) is, however, concerned with situations in which the accuracy of the basic data is predetermined. Because they

acquire a large part of their primary data themselves, forest management agencies are in the fortunate position of being able to chain backwards through the various GIS operations in order to define the accuracy of primary data acquisition needed to achieve the desired quality of the information product. In this way, the methods and costs of spatial data acquisition can be better matched to the value of the resource being mapped.

### ACKNOWLEDGEMENT

This paper is based on information about spatial data acquisition kindly supplied by forestry agencies in many parts of the world.

### REFERENCES

- Aspinall, R.J., Miller, D.R. and Birnie, R.V. 1993. Geographical information systems for rural land use planning. *Applied Geography*, 13:54-66.
- Corrie, G., Reedijk, W. and Lohia, K. 1994. A GIS to improve the competitive advantage of TimberWest Forest Limited. *Proc. GIS'94, Vancouver:83-88.*
- Gemmell, F., Goodenough, D.G., Fung, K. and Kushigbor, C. 1991. Resource spatial and attribute information extraction from remotely sensed data. *Proc. Spatial Data 2000, Oxford:221-231.*
- Griess, O. 1989. Die Forsteinrichtung, ein klassisches geographisches Informationssystem. *Proc. Symposium on New Technologies in Forest Management Planning, IUFRO S4.04, Luxemburg:8pp.*
- Heuvelink, G.B.M. 1993 *Error Propagation in Quantitative Spatial Modelling: Applications in Geographical Information Systems. Nederlandse Geografische Studies 163, KNAG, Utrecht:151pp.*
- Keefer, B.J. 1994. Impact of spatial accuracy on business decisions: a forest management perspective. *Proc. International Symposium on the Spatial Accuracy of Natural Resource Data Bases, Williamsburg:18-23.*
- Kennedy, P.J., Päivinen, and Roihuvuo, L. (eds.) 1994. *Proc. International Workshop on Designing a System of Nomenclature for European Forest Mapping. European Commission, Luxemburg:455pp.*
- Leckie, D.G. and Gillis, M.D. 1995. Forest inventory in Canada with emphasis on map production. *Forestry Chronicle*, 71(1):74-88.
- Moser, W. 1970. Wie genau muß eine Forstkarte sein? *Allgemeine Forstzeitung*, 81(8):205-207.
- Prisley, S. P. 1994. Why natural resource information managers must be concerned about spatial accuracy. *Proc. International Symposium on the Spatial Accuracy of Natural Resource Data Bases, Williamsburg:24-34.*
- Pröbsting, T. 1994. Einsatz von Luftbildern in der Forstwirtschaft in Europa - Ein Überblick. In T. Pröbsting (ed.) *Photogrammetrie & Forst. Albert-Ludwigs-Universität, Freiburg:277-296.*

- Wander, R. 1981. Die Forstvermessung in Landesteil Baden. Allgemeine Vermessungsnachrichten 8/9:329-338.
- Weir, M.J.C. 1981. Simple plotting instruments for resource mapping. Proc. International Conference on Matching Remote Sensing Technologies and their Applications, London:223-232.

### **BIOGRAPHICAL SKETCH**

Michael Weir is a lecturer at the International Institute for Aerospace Survey and Earth Sciences (ITC), Enschede, The Netherlands. He is mainly responsible for teaching surveying, photogrammetry and GIS to foresters and other resource specialists from developing countries.

# Attribute and Positional Accuracy Assessment of the Murray Darling Basin Project, Australia

Fitzgerald, R.W.<sup>1</sup>, Ritman, K.T. & Lewis, A.

---

**Abstract:** The Murray Darling Basin comprises a land area of 1,058,000 km<sup>2</sup> covering a substantial portion of SE Australia and encompassing Australia's largest river system. A basin wide woody vegetation dataset has been assembled from LANDSAT TM imagery supplemented with aerial photography at a nominal scale of 1:100,000.

The attribute accuracy assessment method is based on a multi-stage systematic sample design. A rectangular grid is placed over the entire Murray Darling Basin dataset and at each primary sample point, a secondary grid is created formed from a square contiguous set of aerial photos.

A half tone, grey scale transparency covering approximately 10km<sup>2</sup> is generated from satellite digital imagery (LANDSAT TM) for each primary grid point at a contact scale matching the available air photos. Overprinted on this transparency are the secondary grid points as patch size sampling frames. The air photos are then visually co-registered underneath the transparency with Landcover features from the LANDSAT TM image. Attribute data (presence or absence of woody vegetation) is collected directly from the air photos for each secondary grid point at 4 different patch sizes.

Positional accuracy is assessed by recording 40 or more ground control points from the 1:100,000 map sheet containing the primary grid point. These Eastings and Northings are compared to their position in the LANDSAT TM image.

The outcome of the accuracy assessment is a Basin wide attribute and positional accuracy statement and spatial variability contour maps of attribute and positional accuracy as GIS overlays.

## INTRODUCTION

This paper discusses the attribute and positional accuracy assessment

---

<sup>1</sup> *Infoplus, Po Box 125, Queanbeyan, NSW 2620, Australia. Fax: +61 (0)6 299 1331, E-mail: bfitzger@pcug.org.au*

methodology developed for the Murray Darling Basin Project<sup>2</sup> (MDBP). The Murray Darling Basin (MDB) covers an area exceeding 440 x 1:100,000 map sheets with a total area of 1,058,000 km<sup>2</sup>. The diverse range of vegetation and land forms creates challenges in terms of methodology and logistics.

The initial focus of the accuracy assessment method include three stratification hypotheses. The first is that the local geographic variations in the classification accuracy is a function of vegetation type, terrain and substrate. Second, the smaller the patch of woody vegetation, the lower the attribute accuracy. Thirdly, the overall accuracy percentage is proportional to the percentage of vegetation cover. If the vegetation cover is patchy, then the ratio between the length of a boundary around a patch polygon and its enclosed area is of interest. This ratio was examined by Crapper et.al. (1986).

The positional and accuracy assessment methodology will be:

- a. Statistically sound, practical, inexpensive to implement, easily understood by project staff and portable to MDBP's GIS;
- b. Produce scalable attribute and positional accuracy assessments of the woody/ non woody vegetation dataset as;
  - i. Attribute accuracy assessment statistics (error matrix, user and producer accuracies and Kappa statistic) at different spatial scales;
  - ii. Spatial variability maps of attribute and positional accuracy for the entire MDBP.

A brief examination of the MDB woody/non-woody GIS layer provided an insight into data quality and processing standards. The details of these standards including lineage are outlined in the Draft specification (Ritman, 1995).

The Victorian & South Australian groups have examined attribute accuracy assessment methodologies. The work of Czaplewski et.al. (1992) proved to be the most substantial and well documented attribute accuracy assessment methodology available. They used aerial photo interpretation as their pseudo ground truth and a systematic sample. Tadrowski et.al. (1990) in South Australia investigated supervised classification aided by manual photo interpretation and a stratified simple random sample for attribute accuracy assessment.

The production of the digital woody vegetation dataset at a nominal Basin-wide scale of 1:100,000 was achieved by a two stage process (Ritman, 1995). Woody vegetation is defined as any perennial vegetation having a height exceeding 2m and a density greater than 20% crown cover (McDonald et.al., 1990).

The first stage was the digital classification from LANDSAT TM imagery of a template of only woody vegetation. This method is based on that of Gilbee & Goodson (1992) and comprises an unsupervised 100 ISO class cluster analysis of

---

<sup>2</sup> This paper is a product of a consulting project titled "Attribute Accuracy Assessment for Project M305", DLWC, September 1995, Murray Darling Basin Project M305.

LANDSAT TM imagery, followed by manual aggregation based on field input, aerial photos and ancillary data. In addition a filter (an ARCINFO AML process) designed to remove patches of cells diagonally and orthogonally connected, created by Dr. K. Ritman, was passed over the resulting woody vegetation layer to remove unconnected vegetation patches less than 0.25 ha.. The resultant woody dataset is in raster format with 25 x 25m pixels.

The next stage was either a manual or digital classification of vegetation structural elements such as genus, density class and growth form. Only the woody vegetation layer is subject to the attribute and positional accuracy methodology described in this paper.

### **A BRIEF REVIEW OF THE LITERATURE AND PREVIOUS STUDIES.**

Attribute classification accuracy is usually assessed by constructing a contingency table of a classified map versus ground truth or reference data (Congalton, 1991; Veregin, 1989). The resulting error or confusion matrix  $C$  is a  $k \times k$  matrix where  $k$  is the number of discrete classes in the classification scheme. In the case of the MDB woody/non-woody dataset,  $k = 2$ .

The most commonly used index of attribute classification accuracy is the overall accuracy percent (OA%). Confidence limits for the OA% can be constructed easily from either the binomial distribution or the normal approximation to the binomial distribution. Van Genderen et.al. (1978) logically extends the use of the binomial confidence limits to estimate sample sizes given expected classification accuracy and confidence levels.

The OA% is a simple index of classification accuracy which has its limitations. The OA% can't differentiate between errors of omission and commission. Also it can not reliably be used to compare the performance of different error matrices with different sample sizes as well as not being able to account for correct classification by chance alone.

One of the best methods developed to overcome these limitations is the Kappa statistic discussed extensively by Congalton (1991) and Fitzgerald & Lees (1994a). It statistically quantifies the level of agreement and has been shown to give a less biased estimate of classification accuracy than the OA% (Rosenfield & Fitzpatrick-Lins, 1986).

The effects of sampling schemes on classification accuracies especially when viewed across the spectral, spatial, environmental, taxonomical and temporal domains can induce substantial bias into classification. Congalton (1988, 1991) compares the relative effects of five sampling schemes including random and stratified random sampling on classification accuracy. Franklin & Hiernaux (1991) discuss the effect of sampling schemes on woody vegetation classification while Fitzgerald & Lees (1994b) discuss scale and its relationship to floristic structure.

High accuracy surveying standards exist for assessing the accuracy of topographic maps in all three dimensions. The state and national mapping agencies are responsible for the surveying standards and integrity of the map base. The implicit assumption made in this study is that this *map base is generally correct*<sup>3</sup>.

Positional accuracy quantifies the accuracy of feature locations after various image processing and GIS transformations have been applied. A number of tests are available to assess positional accuracy including deductive estimates, internal evidence checks, comparisons to source documents and reference to independent sources of higher accuracy (Veregin, 1989). The latter is the most desirable and the one used in this study. The independent source is the AUSLIG 1:100,000 map base series.

Spatial variability maps of both attribute and positional accuracy will be produced for the MDB from this accuracy assessment methodology. The 12 x 18 primary systematic grid (described below) will contain the derived data values of OA% and Kappa (attribute accuracy) and 2D-RMS and CEP (positional accuracy). These gridded values will then be interpolated to a surface from which a contour map (isometric lines) of attribute and positional accuracy will be produced.

The accuracy of this interpolation is dependant on the number and spatial distribution of the observed sample values. Systematic sampling (aligned or unaligned) is the best of all the sampling techniques tested to minimise the effects of spatial distribution on contouring (Veregin, 1989).

## SAMPLE DESIGN

The constraints on the sample design for the attribute and positional accuracy assessment for the MDBP were as follows:

- a. The extent of the MDB (1,058,000km<sup>2</sup>) precludes field checking as the major source of ground truth. Aerial photo interpretation is the only practical means in this case;
- b. The design must be practical, simple and expedient to implement. Specifically the handling of maps, air photos, air photo run maps along with satellite imagery should be handled as efficiently as possible with the minimum number of each being accessed as possible;
- c. Attribute accuracy must be assessed at a minimum patch size of 1 hectare across the entire basin. The patch size scale effect on classification accuracy should be assessed if possible at scales of 0.25, 1.00, 4.00 and 9.00 ha.;

---

<sup>3</sup> This assumption proved incorrect. Field experience of project staff demonstrated that the map base is not always reliable. Combined with budgetary restrictions, the positional accuracy component of the methodology was cancelled in the implementation phase.

- d. The results of the accuracy assessment must be statistically defensible and easily interpreted by the end user community.

With these constraints in mind a review of the literature suggested that the four best contenders for the sample design were: Simple random sampling (SRS); Stratified simple random sampling (STSRs); Cluster sampling; Systematic sampling (SYS).

Simple random sampling has the advantage of being the easiest to construct. Implementation especially in the field can be problematic. Statistically, the estimates produced are easily produced and are robust. For the purposes of the MDBP, SRS was considered impractical to implement on such a large scale dataset.

Stratified SRS is the most often used design as judged from literature. It is statistically more efficient than SRS (Cochran, 1977), can produce less biased results than SRS and is a little easier to implement (Congalton, 1988; Janssen & van der Waal, 1994; Van Genderen et.al., 1978). Constructing a stratified SRS is more complex than a SRS by itself and stills suffers from the problems of access to ground truth sites. The claimed statistical gains over SRS are also very dependent on the spatial autocorrelation structure of the dataset (Congalton, 1988) which is unknown. Stratified SRS was not considered any more practical to implement than SRS for the MDBP.

Cluster sampling is the preferred design when cost of access to the ground truth sites is at a premium. Moisen et.al. (1994) showed that cluster sampling with a fixed cost had a higher relative efficiency than either systematic or simple random sampling. The construction and implementation of cluster sampling are more complex than either stratified or simple random sampling. Once again the practicalities of this design precluded its use in the MDBP.

Systematic samples have as their most attractive feature their ease of construction and implementation. Systematic sampling is particularly suited to spatial problems involving two dimensions or more as noted by Cochran (1977). Cochran (1977) and Congalton (1988) propose that unaligned systematic sampling is superior to aligned or centred systematic sampling.

A large number of authors and studies have used systematic sampling. Goodchild et.al. (1991) utilised a single stage systematic sample of 1347 sites in the CALVEG study. Czaplewski et.al. (1992) used a systematic sample of 363 plots for the Victorian CNR Tree cover project.

The attribute accuracy assessment method recommended for the MDBP is a two stage systematic sample with subsampling units of equal size. The sampling unit is a woody vegetation patch. The sample design has the practical advantage of simplifying the acquisition and handling of the aerial photos used to acquire the pseudo ground truth values for the attribute accuracy assessment.

In the first sampling stage, a 12 x 18 rectangular primary sampling grid (216 primary grid points) is placed over the MDB dataset. This grid is oriented 40° from North to incorporate linear trends in Landcover categories.

At each primary sample point, a 7x7 secondary grid (49 secondary grid points) is created orientated to the AMG grid. Thus the number of secondary sample points will be:  $12 \times 18 \times 7 \times 7 = 10,584$ . Due to the irregular outline of the MDB, approximately 10% of the primary grid point will fall outside of the Basin and will be excluded. Thus the final sample size will be less than 10,584, possibly around 9,500. The sampling unit is a woody vegetation patch.

This two stage design has the added advantage of being scalable. The grid cell size at either sample stage can be varied to reflect desired confidence limits, budget and time constraints<sup>4</sup>. The presence/ absence of woody vegetation at the 4 patch sizes (0.25, 1, 4 and 9 ha.) will be collected at each Secondary grid point. This data will be used to investigate the effect of patch size on classification accuracy.

To collect the attribute information, a half tone, grey scale transparency covering approximately 100km<sup>2</sup> is generated from satellite digital imagery (LANDSAT TM) for each primary grid point at a contact scale matching the available air photos. Overprinted on this transparency are the secondary grid points as patch size sampling frames. The air photos are then visually coregistered underneath the transparency with Landcover features from the LANDSAT TM image (MDBC, 1995).

Attribute data (presence or absence of woody vegetation) is collected directly from the air photos for each secondary grid point at 4 different patch sizes. This information forms the pseudo ground truth values for comparison to the LANDSAT TM derived vegetation layer (woody/ non woody). The attribute data for each of the 49 secondary grid points is compiled and crosstabulated with the classified LANDSAT TM woody vegetation values. An OA%, user and producers accuracy along with the Kappa statistic are derived from this error matrix to assess the attribute accuracy at each primary grid point. These estimates are statistically valid at the secondary grid point level.

Positional accuracy is assessed by recording 40 or more ground control points (GCPs) from the 1:100000 map sheet containing each primary grid point. These Eastings and Northings are compared to their position in the LANDSAT TM image. From these GCPs, the mean and standard deviation of the absolute differences along with the 2D-RMS and Circular Error probability (CEP) radius are computed.

The specified positional accuracy of the woody vegetation dataset at 1:100,000 scale is  $90\% \leq 50$  metres. The results from the 3 spot checks indicate that all the mean 2D-RMS values are  $> 50$ m and the 85% CEP are 2 to 3 times the 50m specification.

---

<sup>4</sup> Budgetary constraints during the implementation of the accuracy assessment methodology dictated that the sampling grids be reduced to 10 x 15 primary and 6 x 6 secondary points giving a total of 5,400 sampling points.

## SAMPLE SIZE DETERMINATION

The sampling unit for the attribute accuracy assessment is a patch of woody vegetation. The minimum Basin wide patch size is 1 ha. The total number of 1 ha. patches is  $N = 105,800,000$  and at the minimum mapping unit size of 0.25 ha.,  $N = 423,200,000$ .

One of the statistical problems faced in the sample design is that there are very few accepted guidelines for determining sample sizes in spatial analysis. Many well respected authors and studies pluck a sample size from the ether with naught justification. Goodchild et.al. (1991) uses a sample size of 1347 sites systematically sampled from 56,973 sites and does not described how they arrived at this figure.

The beginnings of the statistical definition of sample size in spatial analysis is seen in the early work of Van Genderen et.al. (1978). They outline the use of the binomial distribution as a means for deciding the sample size based on a confidence requirement (95%) and a specified minimum classification error in the population of 85% correct. This corresponds with the US Geological Survey Circular 671 operational job specification. Rosenfield & Melly (1980) and Veregin (1989) provide a similar rationale based on the confidence interval for proportions, again based on the binomial distribution.

At the secondary grid level, the proposed sample design takes an area of 10x10km. This corresponds to 10,000 x 1 ha. patches. Thus based on the binomial theory, a sample of 60 one hectare patches from 10,000 one hectare patches, a sample fraction of 0.6%, should provide an unbiased estimate of the population (here a 10 x 10km area) classification accuracy.

However, at the Basin wide level, 60 one hectare sample points amongst 105,800,000 (a sample fraction of 0.0001%) is a tad small! Theoretically it is justifiable. Typical telephone polling of the entire Australian population has sample fractions of 0.01%, 100 times that of the above!

Unfortunately, the Remote Sensing & GIS literature gives little guidance as to what constitutes an acceptable sample size. Discussions with Dr. Ray Czaplewski and Gretchen Moisen confirmed this view. Congalton (1991) offers a *rule of thumb* of 100 sample sites per classification category. Czaplewski & Catts (1992) recommend sample sizes of 500 to 1000 based on simulation studies.

The decision on the final sample sizes was constrained by 3 factors. The number of primary sampling points needed to be as large as practicable to minimise the contour interpolation error. The secondary grid sample size of 60 (95% correct classification with 95% confidence) based on the binomial distribution seemed reasonable based on the literature and discussions with colleagues. The sample size had to accommodate the patch size sampling frames on the transparent overlay.

The compromise decided on was a primary grid of 12x18 (216 points) and a

secondary grid of 7x7 (49 points) giving a total sample size of 10,584 patches. The secondary grid size is statistically defensible and the primary grid size gives sufficient control points for the contour interpolation.

### **ANALYSIS OF THE ATTRIBUTE AND POSITIONAL ACCURACY.**

The Positional accuracy assessment will be based on a summary table of the absolute differences between the map and digital image coordinates (E & N) of the GCPs recorded at the secondary grid level for each primary grid point. The GCP outliers (extreme differences i.e. > 1km), will be identified, documented and removed from the analysis. Summary statistics (means, standard deviations, maximum & minimum) for each primary grid point will be aggregated into a Basin wide positional accuracy report. The 2D-RMS and Circular error probability radius (CEP) for each primary grid point will be created from the raw GCP differences discussed above. A contour plot of the 2D-RMS from each primary grid point will be produced. This contour map illustrates the spatial variability of positional accuracy across the MDB.

The Attribute accuracy assessment follows the trend in most of the literature in utilising an analysis of the error matrix. For each Primary grid point and for each patch size, an error matrix of the pseudo ground truth values of woody/ non woody derived from the air photo interpretation versus the values derived from the digital imagery is constructed.

From this error matrix (for each patch size) the Overall Accuracy % (OA%), Kappa statistic and user and producer accuracies are computed for each primary grid point and for each patch size. They can be reported separately as the secondary level sample size is sufficient to make them statistically valid. The contents of the Basin wide attribute accuracy summary table is then contour mapped to produce the spatial variability maps of Attribute accuracy.

### **CONCLUSIONS**

The attribute accuracy assessment project developed and trialed a methodology for assessing the attribute and positional mapping accuracy of the woody vegetation layer (woody/ non-woody) within the MDB dataset. The flexibility of two stage systematic sampling, its simplicity of implementation, the large number of regular grid points distributed across the basin at different sampling levels and the potential flexibility for post stratification and interpolation were the deciding factor in choosing it for the attribute accuracy assessment for the MDBP.

### **ACKNOWLEDGEMENTS**

I'm indebted to the practical experience of Adam Choma of CNR, Peter Knock

of DLWC, Bathurst and Graeme Dudgeon (NSW Dept. Agriculture, Orange). Mrs. Kim Smith my research assistant persevered with the GCPs. On the statistical front, Dr. Ray Czaplewski of the US Forestry Service, Fort Collins Colorado USA proved an invaluable sounding board. Gretchen Moisen of the US Dept. Of Agriculture, Ogden Utah USA provided insights into accuracy assessment on large scale datasets. Dr. Brian Lees (Australian National University, Geography Dept.) provided early thoughts.

## REFERENCES

- Cochran, W.G., 1977, *Sampling Techniques* (3rd ed.), John Wiley & Sons, NY
- Congalton, R.G., 1988, Comparison of sampling schemes used in generating error matrices for assessing the accuracy of maps generated from remotely sensed data., *Photogrammetric Engineering and Remote Sensing*, v 54, n 5 May 1988, p 593-600
- Congalton R.G., 1991, A review of assessing the accuracy of classifications of remotely sensed data., *Remote sensing of Environment*, v 37, n 1, p 35-46
- Crapper, P.F., Walker, P.A., Nanninga, P.M., 1986, Theoretical prediction of the effect of aggregation on grid cell data sets., *Ge-processing*, 3, p155-166.
- Czaplewski, R. & Catts, G.P., 1992, Calibration of Remotely Sensed proportion or area estimates for misclassification error. *Remote Sensing of Environment*, v 39, p 29-43
- Czaplewski, R., Goodson, P., Gilbee, A., Razier, P., Choma, A., 1992, *Accuracy assessment of Remotely Sensed thematic maps*. Project Brief, Dept Conservation and Environment, Victoria, Australia, September 29, 1992.
- Fitzgerald, R.W. & Lees, B.G., 1994a, Assessing the classification accuracy of multisource remote sensing data., *Remote Sensing of Environment*, 47, n 3, 362-368
- Fitzgerald, R.W., & Lees, B.G., 1994b, *Spatial Context and scale relationships in raster data for thematic mapping in natural systems*, Spatial Data Handling Conference, Edinburgh, Scotland, Sept 1994. Published by Taylor & Francis, UK (in press).
- Franklin, J., Hiernaux, P.H.Y., 1991, Estimating Foliage and Woody Biomass in Sahelian and Sudanian Woodlands Using a Remote-Sensing Model., *International Journal of Remote Sensing*, v 12, n 6, p 1387-1404
- Goodchild, M.F., Davis, F.W., Painho, M., Stoms, D.M., 1991, *The use of vegetation maps in geographic information systems for assessing conifer lands in California*. National Centre for Geographic Information and analysis, Dept Geography, Uni California. Report 91-2B NCGIA
- Gilbee, A. & Goodson, P., 1992, *Mapping tree cover across Victoria using Thematic Mapper digital data and a GIS.*, Proc. 6th. Australian Remote

- Sensing Conference, Wellington, NZ, Nov 1992.
- Janssen, L.L.F. & van der Wel, F.J.M., 1994, Accuracy assessment of satellite derived land-cover data: a review, *Photogrammetric Engineering & Remote Sensing*, v 60, n 4, April, p 419-426
- McDonald, R.C., Isbell, R.F., Speight, J.G., Walker, J. & Hopkins, M.S., 1990, *Australian Soil and Land Survey; Field Handbook*, ed. 2, Inkata Press, Melbourne, Australia.
- Moisen, G.G., Edwards, T.C. Jnr, Cutler, D.R., 1994, *Spatial sampling to assess classification accuracy of Remotely Sensed data*. Environmental Information Management and Analysis: Ecosystem to Global Scales Michener, Brunt and Stafford (eds). p 159-176
- Murray Darling Basin Commission, 1995, *Recipe for the Attribute and Positional Accuracy Assessment of the Murray Darling Basin Project.*, internal working document, by R W Fitzgerald, Infoplus for DLWC Land Information Centre, Bathurst, Australia.
- Ritman, K.T., 1995, *Structural vegetation data; A specifications for the Murray Darling Basin Project M305*. DLWC, Land Information Centre, Bathurst, Australia
- Rosenfield, G.H. & Fitzpatrick-Lins, K., 1986, A coefficient of agreement as a measure of thematic classification accuracy., *Photogrammetric Engineering and Remote Sensing*, v 52, n 2, p 223-227.
- Rosenfield, G.H. & Melly, M.L., 1980, Applications of statistics to thematic mapping, *Photogrammetric Engineering and Remote Sensing*, v 46, n 10, October, p 1287-1294
- Tadowski, T., Hart, D.G.D., Schepp, K., 1990, A study of the possibilities and accuracies of 1:50 000 vegetation mapping using Remotely Sensed data. *8th Australian Inst Cartographers Conference*, Darwin, Australia, May 1990.
- Van Genderen, J.L., Lock, B.F. & Vass, P.A., 1978, Remote Sensing: Statistical testing of thematic map accuracy. *Remote Sensing of the Environment*, v 7, p 3-14.
- Veregin, H., 1989, *A taxonomy of error in spatial databases*. National Centre of Geographic Information and Analysis, Technical Report 89-12, December 1989.

# A New Method for Evaluating Positional Map Accuracy

Michal Lodin<sup>1</sup> and David Skea<sup>2</sup>.

---

**Abstract.**-Typically, positional map accuracy is determined from a small number of 'well defined points'. If a sufficient number of such points exist, it is assumed that they are evenly distributed over the map, and represent the overall map accuracy. In this study, a method of direct examination of the geometry representing point, linear, and aerial features is described. Pattern recognition procedures are used to automatically extract and match all identifiable nodes, arcs, and polygons of an accurate benchmark map, and a test map. Different measures of "similarity" are computed for corresponding map elements, - cut blocks and logging roads. Due to the automated approach, large datasets containing virtually all matching features are compiled. Thus, distribution, correlation, and spatial variation: tests, typically not done in standard accuracy assessments, can be carried out. Various mapping methods, operationally used in maintaining the inventory map database of the British Columbia Forest Service were tested using the developed methodology.

---

## INTRODUCTION

The British Columbia Forest Service manages well over six thousand 1:20 000 mapsheets covering an area of approximately 83 million hectares. Forest openings resulting from timber harvesting and road construction are updated on a two year cycle. The map update process is carried out both by Forest Service staff and through contractors. The objective of this study was to develop and test a suitable methodology to evaluate positional accuracy of map update procedures used in the Forest Inventory program. Maps obtained by operationally available methods were compared to a highly accurate benchmark through a set of objective procedures which provide a quantitative measure of their mutual "dissimilarity".

Positional mapping accuracy standards, as adopted by most modern mapping programs, state that a percentage of all well-defined planimetric features are located within some minimum distance of their geographic position with reference to a prescribed datum (Committee for Standards and Specifications, 1985). To test

---

<sup>1</sup> *Remote Sensing Specialist, Ministry of Forests, Victoria, British Columbia, Canada*

<sup>2</sup> *Systems Analyst, Ministry of Environment, Lands and Parks, Victoria, British Columbia, Canada*

if a given map meets a stated standard, the location of a number of well-defined points are sampled and compared with their true locations, as determined by some highly accurate positioning method (e.g., differential GPS). The differences between the two positions are then computed and summary statistics such as RMS and CSE (Circular Map Error) are computed. An outline of this methodology can be found in Maling (1989) or Merchant (1987). Another example, applied to forest mapping, is given in DMG (1991). Unbiased selection of well defined points is another problem. Dozier and Strahler (1983) have suggested that all possible well defined points be extracted from the map and then from these a random selection of these points be used as the sample. Lack of resources to collect sufficient ground information is common (Fenstermaker, 1991).

## METHODOLOGY

Both test sites were mapped using the most accurate method available. The resulting benchmark map is of a higher order of accuracy than that produced by the tested methodologies. The same site is then remapped using the test methodologies, and these new maps are compared to the reference map. From the observed differences, a statistical picture of the similarity of the new map compared to the reference map is derived. All tests are empirical, thus no effort has been made to separate different error components. In terms of GIS error taxonomy (Collins and Smith, 1994), most errors described here can be assumed to be category  $E_{1,2,5,6,P,S,E}$  (systematic and random positional errors due to data collection, input, output and interpretation). A diagram showing the developed procedure is given in figure 1. If the reference map is considered error-free, this statistical picture would reflect the accuracy of the new maps.

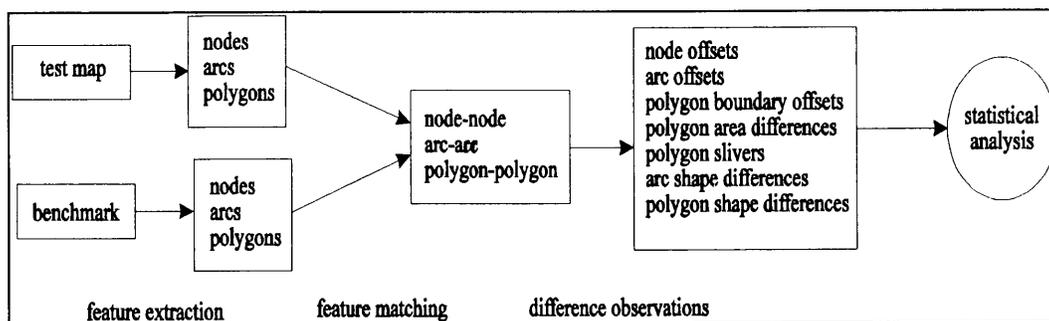


Figure 1.--Process diagram for automated positional map accuracy assessment.

### Benchmark map compilation

Two test sites were selected, both with significant relief and both containing a large number of features to be mapped. High accuracy (as compared to the tested

methods) benchmark maps were compiled for both sites using GPS ground control points and analytical plotter instrument. Positional accuracy of one of the benchmarks was evaluated using seventeen 'well defined points' to be within RMS 7.35 meters ( $X, Y$ ) from their GPS field surveyed positions.

### Feature extraction

Benchmark and test map were compiled in an unstructured format (noded and cleaned result of 'spaghetti digitizing'). Basic geometric objects: a) polygons representing forest openings, b) arcs constituting roads and partial polygon boundaries, and c) nodes identifying road-road and road-forest opening intersections were formed. Subsequently, an internal data structure, the kd tree was constructed, (Ooi, 1990). Each node of the tree contains one point and divides its subdivision of the plane into two parts (Fig. 2, 3). End points of each line segment contain links defining the line segment and the left-to-right sorting of the nodes. Each time an end point is found to lie within a given tolerance of an existing tree node, the order of the tree node is increased. Hence, nodes of order one represent line-ends not connected to another feature. Nodes of order two are simple two-line connections. Nodes of order three or higher represent line intersections. The built segment tree is scanned to extract nodes arcs and polygons.

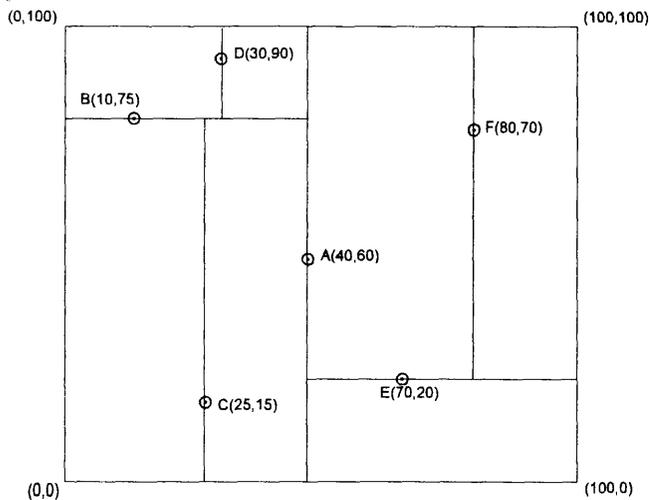


Figure 2.--Points on the plane.

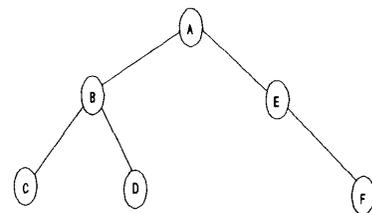


Figure 3.--The resulting kd-tree.

### Feature matching

Automated feature extraction, matching, and differencing procedures have been developed on the MicroStation platform as MDL commands. The matching process was partially hampered by the interpretive nature of the features involved. Polygon boundaries interpreted from LANDSAT imagery did not, in some cases,

resemble the 'true' extent of that polygon on the benchmark map (Fig. 4) and some logging roads were simply not discernible. Matching of extracted geometric objects is facilitated by computing specific attributes for each object and defining some type of 'distance' measure (see Haralick and Shapiro, 1992). An example is given below:

Nodes were matched based on position, number of lines making up the intersection, and relative angle between the lines. For each node on the benchmark map, a range rectangle is defined. Each point on the benchmark which lies within this range rectangle is tested. If the benchmark node has the wrong number of lines then it is rejected. For the remaining nodes, a weighted measure based on the position difference and relative angles is computed and the 'closest' node is accepted as the match.

Arcs and polygons were matched using similar techniques. Polygons matching was close to 100% effective. Node and arc matching is in the order of 80% effective. To correct miss-matches, procedures to interactively view and correct node, arc, and polygon matches have been developed.

Computed difference measures were:

**Node offsets** -  $(\Delta X, \Delta Y)$  in meters,

**Arc and polygon closest point offsets** - Offset of each arc/polygon vertex from the closest vertex of the matching arc/polygon.

**Arc and polygon perpendicular bisectors** - Perpendicular bisector from each line connecting two vertices to the matching line segment of the corresponding arc/polygon line segment.

**Polygon area differences** - Area and perimeter differences between matched polygons.

**Polygon slivers** - Two corresponding boundaries of matching polygons typically cross each other a number of times. These crossings create sliver polygons. For each matched polygon, slivers are extracted and their centroid position, area and perimeter.

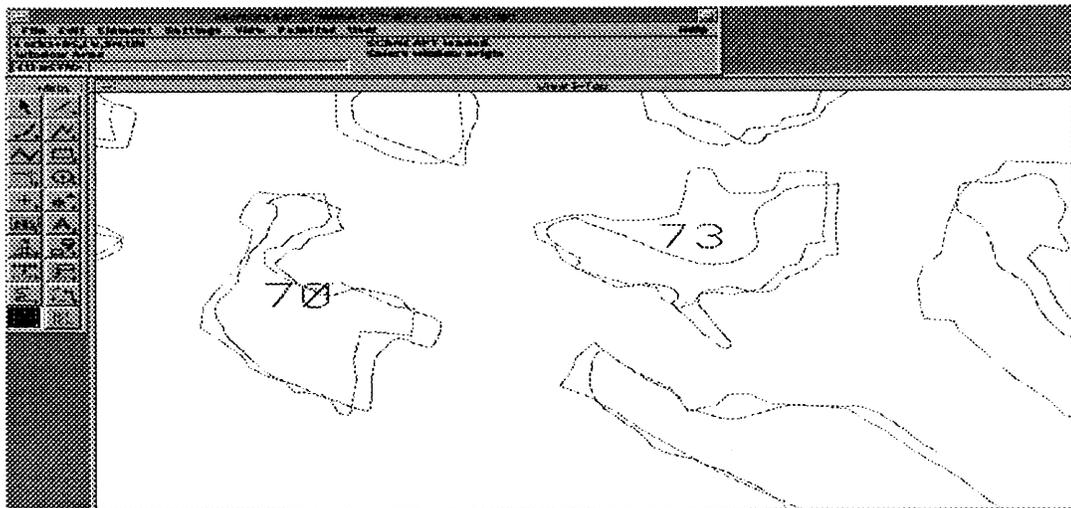


Figure 4.--Example of matched polygons (70) and polygons rejected due to interpretation differences (73).

### Tested methods and compiled datasets

Table 1 shows the fourteen mapping systems tested. The result of all mapping was a digital file in IGDS format. For analog systems, such as PROCOM II, this required an additional manual digitization step (consistent with normal practice).

**Table 1.--Tested methodologies.**

Source data	Format	Mapping method	Contract or in-house production
LANDSAT TM* , 1:1 000 000, 1:500 000	Analogue, transparency*	PROCOM II	Contract
SPOT PAN**, 1:500 000, 1:250 000	Analogue, transparency*	PROCOM II	Contract
LANDSAT TM (Thematic Mapper)	Digital, various levels of geometric corrections	Digital update	In-house
SPOT Panchromatic	Digital	Digital update	In-house
SPOT Multispectral	Digital	Digital update	In-house
Aerial photography, 1:40 000	Analogue	Mono restitution	Contract
Aerial photography, 1:40 000	Scanned @ 853 dpi, 1.2m pixel size	DIAP (Digital Image Analytical Plotter)	Contract
Aerial photography, 1:40 000	Scanned, resampled to 4.8m pixels	DVP (Digital Video Plotter)	In-house

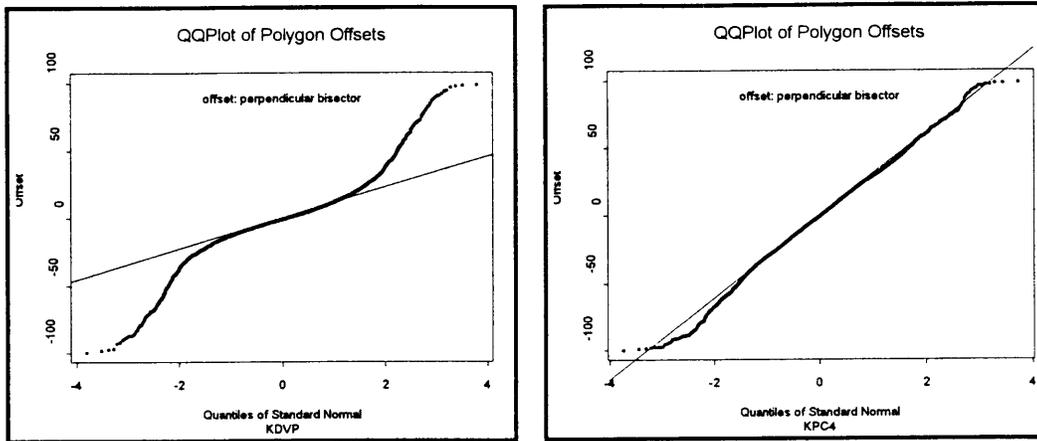
The number of "correctly" mapped and matched features varies from one method to another. Typically, lower order methods, such as optical projection PROCOM II represent the lower bound while higher order softcopy photogrammetric methods (e.g., DIAP) produced results of the upper limit of the range. For some methodologies, the number of matched nodes, ('well defined points' ?) was as low as 10. Large data sets were collected for arc and polygon boundary offsets as well as for polygon slivers. For more detail and test results, see Skea and Lodin (1995).

**Table 2.--Size of compiled datasets.**

Features type matched and evaluated	dataset compiled for one test area
Node offsets road/road and road/polygon	10-300 matched
arc and polygon boundary shortest distance and perpendicular offsets	2,000 - 5,000 observations in each dataset
polygon area, position, shape	50-90 matched
sliver polygons ("in" and "out" relative to benchmark)	400 - 1800 - in each dataset

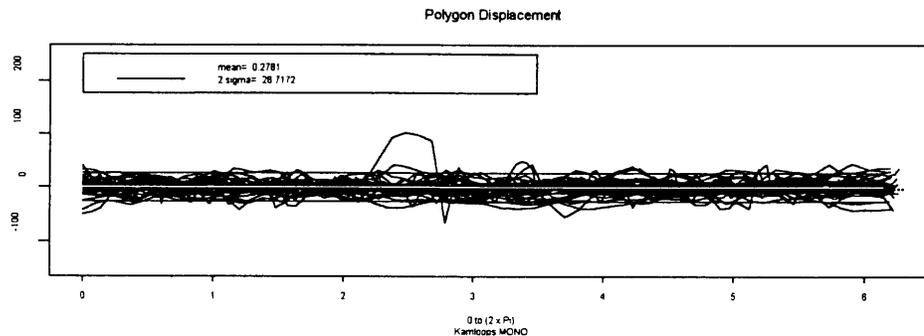
Various distributions of polygon boundary error observations have been reported. Errors associated with manual digitizing (Bolstad *et al* ,1990) were found to conform to a Chi-squared distribution.

Automated observation techniques used in this study produce data sets of sufficient size to directly test for distribution (Fig. 6). These observations suggest, that distribution of offsets cannot be assumed to be normal. Thus, estimates of epsilon bandwidth (Chrisman, 1989), used for polygon and arc boundary error estimates, should be based on distribution free statistics.



**Figure 5.--Quantile-quantile distribution plot of polygon offsets collected by two different methodologies, softcopy photogrammetric device using 1:40 000 photography (L), and an optical projection PROCOM II system using LANDSAT TM 1:500 000 transparency R). Note, in a qq plot, cluster along a straight line indicates normal distribution.**

On the epsilon bandwidth plot (Fig. 6), the median value and epsilon bandwidth or EBW of displacements of polygon/arc boundary are given. The median value shows any skew in the data (ideal data should be centered around zero) and the EBW is computed as:  $+1.4826 \times median \times (y - center)$ . This value corresponds to one sigma (the standard deviation) for Gaussian data but is robust with respect to distribution. The perimeter of each polygon is normalised between zero and  $2\pi$ .



**Figure 6.--Epsilon band plot of polygon displacements for the Mono restitution method.**

## RESULTS

The focus of the statistical analysis was on assessing the form of the data and the relative accuracy, rather than just producing descriptive statistics. This included testing for distribution, computing variance of polygon area differences, and estimating  $\epsilon$ -bandwidth for polygon boundaries and arcs. Standard descriptive measures such as circular map errors were also computed. Results were presented in a comprehensive report (Skea and Lodin, 1995) at different levels of detail in the following categories:

Detection limitations. The number of map features potentially omitted due to data resolution, interpretation limitations or simply due to the “fuzzy” nature of forest openings.

Positional error of intersections (nodes). A statistical measure used for the Provincial base mapping program (Circular Map Standard Error) was found well suited for node accuracy assessment. Surprisingly, positional errors of intersections (nodes) were systematically, for all tested methods, found *not* indicative of the overall map accuracy (Fig. 8).

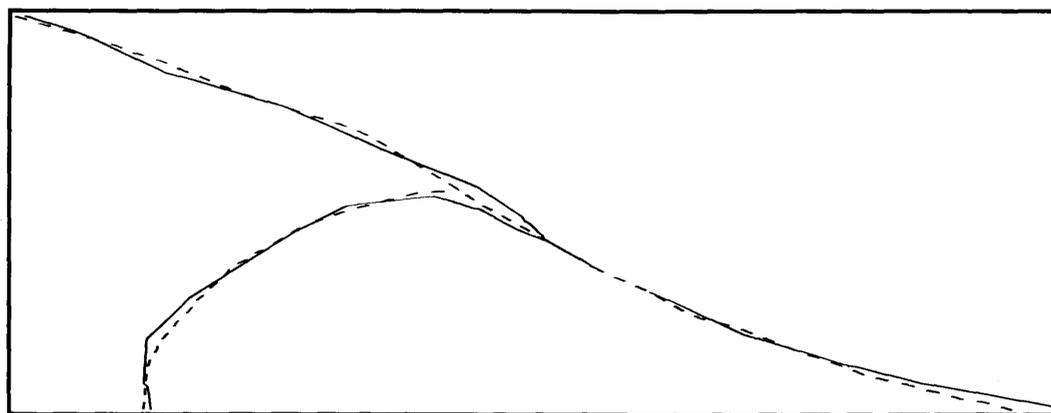


Figure 7.--Example of a node offset which is not representative of overall map accuracy. This 'T' shaped node formed by a road/road intersection captured from a LANDSAT satellite image (dashed line) is compared to the benchmark node which appears skewed. Clearly, the resulting offset is not indicative of the overall accuracy of the map.

Positional error of roads and cutblock boundaries. Epsilon bandwidth was found useful as a means of expressing line (arc and polygon) accuracy. It was used to compare different methodologies, in the context of this study, as it is less dependent on distribution of data compared to other types of statistical measures.

Area error of cutblocks. Overlays of benchmark and test map polygons produce sliver polygons. Areas of sliver polygons in- and out- relative to the benchmark were used as measures of systematic cutblock area under- or over- estimate associated with a given mapping method. Comparing polygon areas only was

found biased as it would favor methods capable of mapping only large cutblock polygons (e.g., LANDSAT).

## CONCLUSIONS

Automated pattern matching techniques can be applied successfully to the problem of positional map accuracy assessment. Their use leads to large and more diverse sets of observations on which to judge map accuracy. Most importantly, positional accuracy of different types of map features cannot be assumed the same for a given method. This conclusion may be affected by nature of mapped features in this study (forest openings). Although the use of these observations clearly requires the development of new statistical analysis procedures, they also allow for the development of broader statistical descriptions of the data.

## REFERENCES

- Bolstad, P.V., Gessler, P. and Lillesand, T.M., 1990. Positional Uncertainty in Manually Digitized Map Data. *International Journal of Geographical Information Systems*, 4 (4), 399-412.
- Chrisman, N.R., 1989. Modeling Error in Overlays of Categorical Maps. In, *Accuracy of Spatial Databases*, edited by, Goodchild, M. and Gopal, S., Taylor and Francis.
- Collins, F.C. and Smith, L.J., 1994. Taxonomy for Error in GIS. In, *Proceedings, International Symposium on the Spatial Accuracy of Natural Resource Data Bases*, 1-7. ASPRS.
- Committee for Specifications and Standards. 1985, *Accuracy Specification for Large-Scale Line Maps*. *Photogrammetric Engineering and Remote Sensing*, Vol. 51, No. 2, 195-199.
- DMG (Digital Mapping Group Ltd.), 1991. Project to Evaluate Photogrammetric Accuracy of Several Plotting Instruments to Transfer Forest Inventory Information from Aerial Photographs to TRIM Base Maps. Technical Report, British Columbia Ministry of Forests, Inventory Branch.
- Dozier, J. and Strahler, A.H.. 1983. Ground Investigations in Support of Remote Sensing. Chapter 23, *Manual of Remote Sensing*, second edition, Simonett, D.S. and Ulaby, F.T. editors, American Society of Photogrammetry and Remote Sensing.

- Fenstermaker, L.K., 1991. A Proposed Approach for National to Global Scale Error Assessments. In, Proceedings, GIS/LIS '91, Vol. 1, 293-300.
- Haralick, R.M. and Shapiro, L.G., 1992. Computer and Robot Vision, Volumes I & II. Addison-Wesley Publishing Company.
- Maling, D.H., 1989. Measurements from Maps: Principles and Methods of Cartometry. Pergamon Press.
- Merchant, D.C., 1987. Spatial Accuracy Specification for Large Scale Topographic Maps. Photogrammetric Engineering and Remote Sensing 53 (7): 958-61.
- Ooi, B.C., 1990. Efficient Query Processing in Geographic Information Systems. Goos, G. and Hartmanis, J. (ed.), Lecture Notes in Computer Science, No. 471, Springer-Verlag.
- Skea, D. and Lodin, M., 1995, Computer-Assisted Methods to Assess the Positional Accuracy of Map Overlays. Ministry of Forests, Resources Inventory Branch, Technical report.

### **BIOGRAPHICAL SKETCH**

Michal Lodin holds Masters degrees in Geology from Charles University, Prague and Surveying Engineering from the University of New Brunswick. After 8 years of consulting in Geomatics applications in Canada, U.S. and Europe, he now works in the Inventory Technical Support Section, Resources Inventory Branch, Ministry of Forests.

David Skea's photogrammetric training includes diplomas from the British Columbia Institute of Technology and the International Institute for Aerial Survey and Earth Sciences in Holland. His research interest are in areas of pattern recognition, computer vision and computational geometry. He is employed with Geographic Data B.C. heading a project to develop an automated watershed and stream network mapping system.

# Exploring Spatial Confidence with a Raster-based GIS

Matthew H. Pelkki<sup>1</sup>

---

**Abstract.**-- One of the main input methods into geographic information systems remains digitizing spatial data from paper maps. While newer, automated methods such as global positioning systems promise better accuracy and quality of data for future GIS, digitizers are likely to be part of the data collection technology for some time into the future. When digitizing this data, it is important to remember that a single sample of the representation of reality is being taken. Information related to the spatial confidence of the resulting representation often is not quantified well by the digitizing software. Spatial confidence can be expressed as percentage or degree of certainty that a represented object does indeed exist at that location or within a certain distance of that location. This research found that for map registration error may not be a good indication the required raster resolution to obtain high spatial confidence. It was also found that complex lines tend to increase the amount of error or variance in the resulting digital map.

---

## INTRODUCTION

Obtaining spatial data of high quality is an obvious desire of spatial data users. Positional accuracy is one important component of data quality (Antenucci et al, 1991). Error in spatial data is inherent and difficult to remove. While minimizing error has long been a goal of spatial sciences (Chrisman, 1991), others (Aronoff 1989) suggest that since error cannot be eliminated, it should be managed. Certainly, the first step in managing error is being able to quantify it.

The need to explicitly understand spatial confidence in a GIS is growing as digital data becomes less specialized and more widespread and mainstream. It is well known that digital data is perceived to be of higher quality than analog data due to the appearance of precision and accuracy. As developers of digital data become more and more removed from the end-users, explicit understanding of locational confidence is important to prevent mis-application of data to analyses for which it is not suitable. Paper maps are currently the most common source for geographic data, and while

---

<sup>1</sup> Assistant Professor, University of Kentucky, Department of Forestry, Lexington, KY 40546-0073.

automated methods such as global positioning systems promise better accuracy and quality of data for future GIS, digitizers are likely to be a part of the data collection technology for some time into the future. Historical data, and much natural resource data will be available only on paper maps. In these cases, the map is in fact the only representation of "truth" and any one manual digitization of that map is but one sample of reality. The map scale is often used as an estimate for the accuracy of the digitized data (Fisher, 1991) and one sample of reality is often all that is taken due to time and cost constraints. This limits the information on the quality of the data, as it assumes that the first sample is truly representative of the "truth."

Ideally, spatial confidence metadata should be attached to a raster GIS that states something like, "there is a 95% confidence that any object classed within a raster cell actually exists within that cell or within a radius of X cells from the indicated cell." From a single sample, even assuming that the analog data represents "truth," this is difficult to measure. This research involves exploration of absolute locational accuracy of point and line data that is manually digitized and rasterized to various resolutions under the assumption that the analog medium being digitized is in fact "truth."

## METHODS

The GIS software chosen for this study was EPPL7 (Environmental Programming and Planning Language) release 2.1. This is a relatively simple raster-based GIS created and maintained by the State of Minnesota Land Management Information Center. It is used as a "desktop" GIS in Minnesota, and has users throughout the United States and in 15 other countries. Users include universities, several state natural resource agencies, the National Park Service, and the Fish and Wildlife Service.

Two separate spatial representations were created digitally. One representation was composed of ten randomly located points. The second spatial representation was composed of ten lines, two each having one to five line segments. This was done to simulate a some increasing complexity in lines, but to keep it on a quantifiable scale. Thus, according to Burrough (1986), more complex lines (those with more vertices) should have a greater degree of error associated with them. These digital representations were considered to be "truth," and were printed out on a relatively stable medium (transparency film) for manual digitizing. The scale of these printed images was 1:50,000, and the point and line width printed on the map was 1/100th of an inch, giving the approximate representation width of < 13 m.

Both the point and the line images were digitized ten times under two conditions. The first set of ten samples were collected under a single map registration, and the second set of ten samples were each collected under a different map registration. This was done to see how map registration differences affected positional accuracy. The registration standard deviation of error was recorded in each and every case. In the

EPPL7 manual, it is recommended that the standard deviation be less than the desired resolution for rasterizing the file (LMIC, 1992). All digitizing was done in the center of a 24 x 36 inch Calcomp 3300 digitizing tablet. The room was climate controlled to reduce climate-caused distortion of the transparencies.

Once digitized, the data files were rasterized to various resolutions (5, 10, 20, and 40 meters). The digital "truth" file was also rasterized to the same resolution. Error in absolute position was determined by overlay, and the error was weighted by the distance between any one raster cell and the "truth."

Similar to the study reported by Maffini et al (1989), this study examines positional errors. It examines the performance of a single digitizer operator over repeated trials and does not control for error properties in converting the digital "truth" to transparency medium, errors in the digitizing tablet or equipment, nor control for speed of the operator.

## RESULTS

Table 1 shows the results from ten separate digitizing operations on the ten point locations using the same map registration. The distances for locational error were calculated by counting the number of raster cells the sample point was displaced from the "truth" point and multiplying by the resolution to obtain X and Y positional error. Since under each resolution there were ten points digitized ten times, the five points with the largest positional errors were discarded, and the positional error that would contain 95 points is recorded in the leftmost column of table 1. We can see that as raster resolution gets smaller, the 95% zone of inclusion approaches a value around 20 m.

Point data digitized under different map registrations is shown in table 2. The rasterized data for 40m resolution were omitted, but a trend similar to table 1 appears

**Table 1. Positional error of point data digitized under same registration and rasterized from 5m to 40m resolution.**

Raster cell resolution	Mean Distance Positional Error	Minimum Positional Error	Maximum Positional Error	95/100 Cells lie within:
5 m	15.8 m	5 m	25 m	21.2 m
10 m	16.8 m	0 m	28.3 m	22.3 m
20 m	14.7 m	0 m	28.3 m	28.3 m
30 m	21.6 m	0 m	42.4 m	30 m
40 m	18.1 m	0 m	56.6 m	40 m

1 - Std. Dev. of registration = 3.16 m

**Table 2. Positional error of point data digitized under different registrations and rasterized from 5m to 30m resolution.**

Raster cell resolution	Mean Distance Positional Error	Minimum Positional Error	Maximum Positional Error	95/100 Cells lie within:
5 m	15.2 m	0 m	28.3 m	25.0 m
10 m	15.1 m	0 m	31.6 m	28.3 m
20 m	13.0 m	0 m	28.3 m	28.3 m
30 m	20.8 m	0 m	42.4 m	42.4 m

1 - Std. Dev. of registrations ranged from 3.69m to 13.17m, and averaged 6.93m

to occur. As raster resolution gets smaller, a more accurate measurement of absolute position error is possible, and the maximum value which includes 95% of all points approaches some value.

Determining positional error for line data is a bit more complex. Error for lines includes overshoots and undershoots, as well as locational deviation for portions of the line. A line error index was calculated for each digitized line as it was compared to the “truth” line. The error index weighted errors by the magnitude of the deviation from the “truth” line. The formula used is as follows:

$$LEI = \frac{\sum_{i=1}^N rd_i + Abs[N' - N]r}{Nr}$$

where: LEI = location error index  
 N = number of raster cells in “truth” line  
 N' = number of raster cells in digitized line  
 r = raster cell resolution  
 d<sub>i</sub> = distance between truth line and digitized line for cell I

This index is independent of line distance, but it is not independent of resolution, since with larger raster cell sizes, small deviations between the “truth” line and the digitized line will be impossible to detect. Therefore, the error index should decrease with raster cell size but increase with line complexity. An LEI = 0.50 means that, for any given line length, 50% of that line lies in a incorrect cell location weighted by the magnitude of the error.

Tables 3 and 4 show the LEI value for lines rasterized to various resolutions under the same and multiple map registrations, respectively. The tables also show the maximum deviation in number of cells for the ten sample lines from the “truth” line. From the tables, two relationships appear. The first is that as raster cell size increases,

**Table 3. Line data digitized under the same registration and rasterized from 5m to 40m resolution (# cells max. deviation in parentheses).**

Raster cell resolution	Number of line segments per line				
	1	2	3	4	5
5 m	1.13 (3)	1.57 (4)	1.51 (7)	1.66 (5)	2.19 (6)
10 m	0.64 (2)	0.80 (2)	0.78 (3)	0.83 (3)	1.10 (3)
20 m	0.39 (1)	0.43 (1)	0.47 (1)	0.48 (2)	0.61 (2)
30 m	0.17 (1)	0.32 (1)	0.30 (1)	0.33 (1)	0.46 (1)
40 m	0.11 (1)	0.23 (1)	0.26 (1)	0.23 (1)	0.35 (1)

**Table 4. Line data digitized under different registrations and rasterized from 5m to 40m resolution (# cells max. deviation in parentheses).**

Raster cell resolution	Number of line segments per line				
	1	2	3	4	5
5 m	1.12 (3)	1.91 (6)	1.64 (7)	1.95 (7)	2.15 (8)
10 m	0.57 (2)	0.93 (3)	0.84 (3)	0.98 (3)	1.07 (4)
20 m	0.32 (1)	0.54 (2)	0.47 (2)	0.52 (2)	0.61 (2)
30 m	0.16 (1)	0.38 (1)	0.33 (1)	0.38 (1)	0.45 (1)
40 m	0.12 (1)	0.28 (1)	0.26 (1)	0.29 (1)	0.35 (1)

the LEI index decreases as does the maximum number of cells deviation for the sample lines. The second relationship is that more complex lines have higher LEI values and maximum cell deviation values, indicating that complex lines introduce more error than simple lines. Multiple linear regressions indicated that the both these relationships are significant at the  $\alpha = 0.05$  level of significance. It is also interesting to note that incorporating the map registration error into the regression model did not significantly improve the prediction of error ( $\alpha=0.10$ ).

## CONCLUSIONS

In the EPPL7 GIS, the summed standard deviation reported during map registration appears to be a poor indicator of locational accuracy of the resulting map. Given the map scale, the point and line representation equated to just under 13 m, and so

perhaps a better estimation of adequate raster cell size would be the 13 m plus two standard deviations. This would suggest a cell size of 20 meters.

It appeared that multiple map registrations increased error, but the tests were uncontrolled and so no comments about the significance of any apparent numerical differences can be made.

In line data, error is strongly correlated to line complexity. This makes good intuitive sense, as the more complex the line, the greater the number of vertices are required to represent that line with reasonable accuracy. The correlation between line error and target raster cell size is an indication that larger raster cells do a poor job of calculating distance, and as the raster cell size increases, differences between the digitized line and the "truth" line that are less than the raster cell size are rounded to zero. However, for maps covering large areas, small raster cell size requires a great deal of processing time and storage.

This preliminary work has identified the need to incorporate various map scales into the procedure, particularly larger scale maps that will better allow the testing of registration error on overall locational error. Testing for multiple digitizing personnel, and various levels of operator experience might also prove interesting.

## ACKNOWLEDGMENTS

Special thanks go to Linda Delay, a student worker that performed the digitizing and ran the computer macros to determine the error components. Without her time and effort on data collection, this project would still be in the conceptual stage.

## REFERENCES

- Antenucci, J. C., K. Brown, P. L. Croswell, M. J. Kevany, and H. Archer. 1991. Geographical information systems: A guide to the technology. Van Nostrand Reinhold, New York, NY. 301 p.
- Aronoff, S. 1989. Geographic Information Systems: A management perspective. WDL Publications, Ottawa, Canada. 294 p.
- Burrough, P. A. 1986. Principles of geographical information systems for land resources assessment. Oxford University Press, New York, NY. 194 p.
- Chrisman, N. R. 1991. The error component in spatial data. In Geographical Information Systems, Volume 1: Principles, D. J. Maguire, M. F. Goodchild, and D. W. Rhind, editors. John Wiley and Sons, New York. Pp. 165-174.
- Fisher, P. F. 1991. Spatial data sources and data problems. In Geographical Information Systems, Volume 1: Principles, D. J. Maguire, M. F. Goodchild, and D. W. Rhind, editors. John Wiley and Sons, New York. Pp. 175-189.
- LMIC. 1992. EPPL7 User's Guide, Release 2.0, Tutorial Chapter, page 110. Land Management Information Center, St. Paul, MN.

Maffini, G., M. Arno, and W. Bitterlich. 1989. Observations and comments on the generation and treatment of error in digital GIS data. In Accuracy of spatial databases, M. Goodchild and S. Gopal, editors. Taylor and Francis, Bristol, PA. Pp. 55-68.

### **BIOGRAPHICAL SKETCH**

Matthew H. Pelkki is an Assistant Professor of Forest Management and Economics at The University of Kentucky Department of Forestry. He graduated with a B.S.F. in 1985 from the University of Michigan's School of Natural Resources & Environment, and then earned an M.S. (1988) and Ph.D. (1992) from the University of Minnesota's College of Natural Resources. Matthew has been an assistant professor at the University of Kentucky since 1991 where he teaches courses in timber management, integrated forest resource management, and applications of GIS in natural resources. His research interests include dynamic programming and stand-level optimization, natural resource information system planning and design, and the economics of bioremediation/ecological restoration.

# GPS Located Accuracy Assessment Plots on the Modoc National Forest

Kevin Casey<sup>1</sup>

---

**Abstract.** A random grid of field plots were evaluated in a blind test to measure the accuracy of remotely sensed vegetation polygons on the Modoc National Forest. Field crews gathered integrated inventory data on all vegetation, as well as collecting accuracy assessment data. GPS positions were collected at plot centers and later post-processed to achieve 2-5 meter horizontal accuracy. Plots were not rotated into condition and field crews had no knowledge of the location of mapped polygons or the label attributes of the polygons. This methodology allows for collection of accuracy assessment data at the same time that inventory data is collected, thereby eliminating the need for separate crews and multiple visits to the field.

---

## INTRODUCTION

Forest vegetation inventories are often based upon a polygon map derived from remote sensing tools. In Region 5 of the U.S. Forest Service (California) vegetation maps are produced at ten year intervals using a combination of TM imagery, color aerial photography and field training sites. The inventory plots are located using a randomized grid of points. The standard grid used in the western U.S.A. has points at 5000 meter intervals. A densified grid is used for conifer forest, having points at 2500 meter intervals. The grid of plots is used to gather basic vegetation composition, structure and growth data. From these data a forest inventory can be compiled. In addition, data needed to perform a formal accuracy assessment of the vegetation map can be collected. This methodology saves time and expense by combining the data collection needs for both the inventory and the accuracy assessment into one project.

## MAPPING

The vegetation map for the Modoc National Forest is based upon 1991 TM imagery which has been terrain corrected and georeferenced. Polygons are created using a clustering algorithm and labels are attached to polygons using a combination of supervised and unsupervised classification. The minimum mapping unit is approximately one hectare, or ten pixels. The CALVEG system of vegetation classification is used to label polygons. In addition, for conifer and

---

<sup>1</sup> Project Leader for integrated resource inventories, Region 5, U.S. Forest Service Remote Sensing Lab, Sacramento, California.

hardwood types, a crown model is used to predict stand size class and density class. The mapping generally occurs prior to the inventory field work, but since the inventory grid is entirely independent of the map, field data collection can occur before or after the map is completed.

## **PLOT LOCATION**

### **Office Procedures**

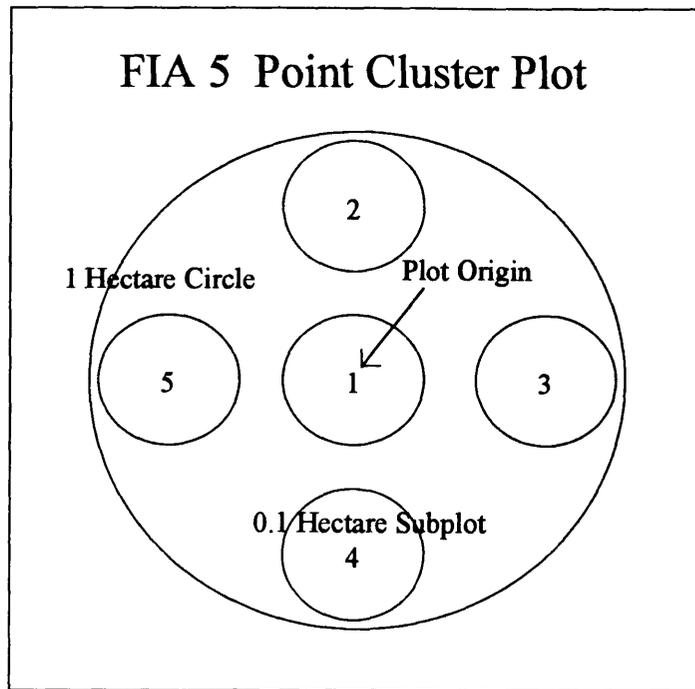
The 5000 meter (or 2500 meter) grid is imported into a GIS in the form of a text file containing UTM coordinate pairs (MOSS format). It is clipped with the forest administrative boundary and then viewed on-screen with a SPOT panchromatic image as a backdrop. Each point of the grid is manually transferred to a color aerial photograph. The photographs are pin-pricked and labeled on the back with a plot number. The plots are also transferred to 1:24,000 scale topographic maps to aid the field crews in locating the plots. UTM coordinates for each plot are printed, also to aid the field crew with the plot location on the ground. No information regarding the vegetation polygon map is put on any of the materials given to the field crew.

### **Field Procedures**

Field crews can use any of the supplied materials to locate the plot center on the ground. The pin-pricked photo is the control, and a tolerance of 20 meters is allowed for error. Crews are equipped with a GPS receiver, and it can be used to navigate to the approximate plot center by entering the UTM coordinates of the plots into the receiver. Final navigation must be done using the aerial photograph.

### **The Cluster Plot**

A five point cluster plot is installed, centered on the pin-pricked location. (See Figure 1.) The total size of the cluster plot is approximately one hectare. At each point of the cluster, vegetation data is measured and entered into a data recorder. A series of circular, nested fixed-area plots are used for collecting vegetation data. Trees are measured individually; shrubs, grasses and forbs are tallied by species and percent cover. The inventory also collects data on snags, down logs and woody debris of various sizes.



**Figure 1. –Plot Configuration**

### **The CALVEG Accuracy Assessment Form**

An accuracy assessment form was developed to capture an independent assessment of the vegetative cover at each sub-plot of the cluster plot. (See Figure 2.) The crew uses a dichotomous key to select the Primary Cover Type and Best CALVEG Type for each of the five subplots. Cover Types are limited to a rather short list; examples being Conifer, Hardwood, Sagebrush shrub, Mixed Chaparral, Herbaceous, etc. Each Cover Type contains many different CALVEG Types. Some examples for the Conifer type are: Mixed Conifer-pine (MP), Mixed Conifer-fir (MF), East-side Ponderosa pine (EP), Western juniper (WJ), and so on.

The plots occasionally fall on an ecotone, or boundary, between two or more distinctive types. Condition codes are used to capture this information. Condition A is used for the vegetation a point #1. If another distinctive type occurs elsewhere on the cluster, it becomes Condition B. For Conifer and Hardwood types, the crew is instructed to evaluate the best size class, and best density class for the stand of trees occupying the plot. In affect, we ask the crew to provide the “best fit” map label for the vegetation at each of the five subplots.

USDA - FOREST SERVICE											REGION 5				
<b>CALVEG ACCURACY ASSESSMENT RECORD</b>															
COMMENTS: _____															
_____															
<b>PLOT/STAND IDENTIFICATION</b>															
NATIONAL FOREST___ PU___ STRATUM___ PLOT NO.____															
DISTRICT___ COMPARTMENT___ STAND NO.____															
1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.
POINT NO.	CONDITION (A,B,C)	PRIMARY COVER TYPE	BEST CALVEG TYPE	NEXT BEST CALVEG TYPE	BEST SIZE CLASS	SECOND BEST SIZE	BEST DENSITY CLASS	SECOND BEST DENSITY	SECONDARY COVER TYPE	BEST CALVEG TYPE	NEXT BEST CALVEG TYPE	BEST SIZE CLASS	SECOND BEST SIZE	BEST DENSITY CLASS	SECOND BEST DENSITY
X	X	X X X	X X	X X	X	X	X	X	X X X	X X	X X	X	X	X	X

Figure 2. --CALVEG data entry form.

### Fuzzy Logic Rules

Often the vegetation on a plot does not fit exactly with the CALVEG descriptions, or has a mosaic of tree sizes and densities occurring on the plot. For these situations “Next Best” ratings can be assigned for any of these parameters. This data provides additional information about trends, and allows for “fuzzy” interpretation of the accuracy data. It is recognized that there is value in determining not only if the map label is right or wrong, but how right or how wrong. The “Next Best” data may indicate that the map label is wrong, but understandably wrong given that the “Next Best” answer matches the label.

### The Plot Diagram

Whenever a “split condition” plot occurs, the crew prepares a Plot Diagram (See Figure 3.) which shows the condition boundary in relationship to the entire cluster plot. A split condition plot can contain two entirely different cover types, such as Conifer and Sage Brush shrub; or two distinctive CALVEG types, such as Lodgepole pine and Red fir; or can result from a distinctive change in tree size or density occurring within a single type. An example might be a dense, undisturbed Red fir stand bordering on a shelterwood unit with widely spaced seed trees.

## GPS OPERATIONS

### Field Procedures

Crews collect GPS data at the plot center (Point #1) only. The specifications require a minimum of 180 positions be logged. The receiver is set up to record 3-D positions only. The elevation mask is set at 15 degrees, the PDOP mask is set at six and the Signal Level mask is set at six. These parameters were selected to

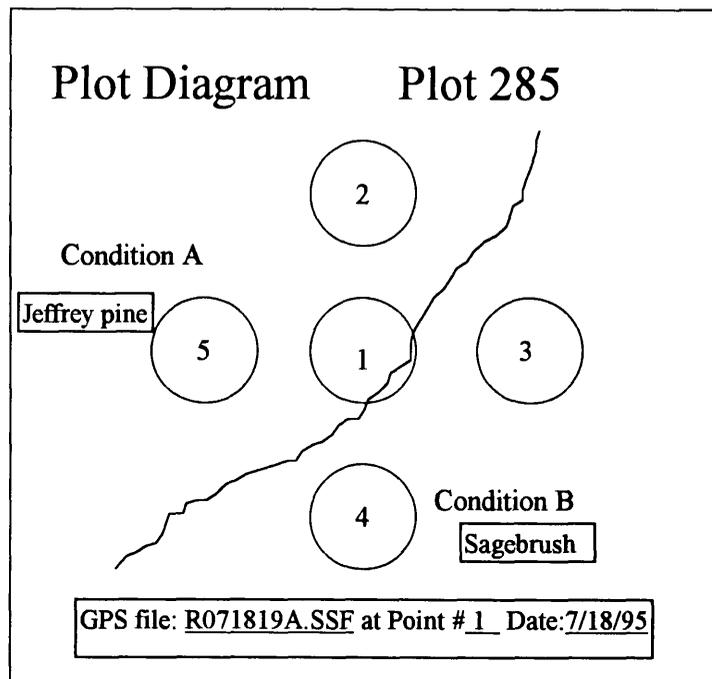


Figure 3. --Example of split condition plot.

produce high quality positions that can be differentially corrected to yield horizontal accuracy of 2-5 meters. Occasionally it is not possible to collect GPS positions at the plot center due to overhanging trees or the terrain. Crews can collect GPS positions at one of the other four subplots, provided they note the location GPS data was collected on the plot diagram.

### Post Processing

The U.S. Forest Service operates three permanent GPS base stations in northern California. A Susanville base station was used to perform differential corrections for the Modoc project. It is located 50 to 150 miles from the plot locations on the Modoc forest. Nearly 100% percent success was achieved in correcting the rover

positions. The mean corrected position for each plot was exported to MOSS format and then brought into a GIS. The actual GPS position of the plot replaces the “predicted” position stored in the database. The difference between the predicted location of the plot center and the GPS located plot center is usually 10-40 meters horizontal distance. Figure 4 illustrates an example of how a plot falls within a vegetation polygon. Note that the “predicted” plot location can result in a plot falling into one polygon, while the GPS position falls into a different polygon. The predicted plot location would suggest that Veg Type B would best match the vegetation data collected in the field. However, the GPS plot location falls primarily into Veg Type A. Using the predicted plot location for an accuracy assessment would yield incongruous results.

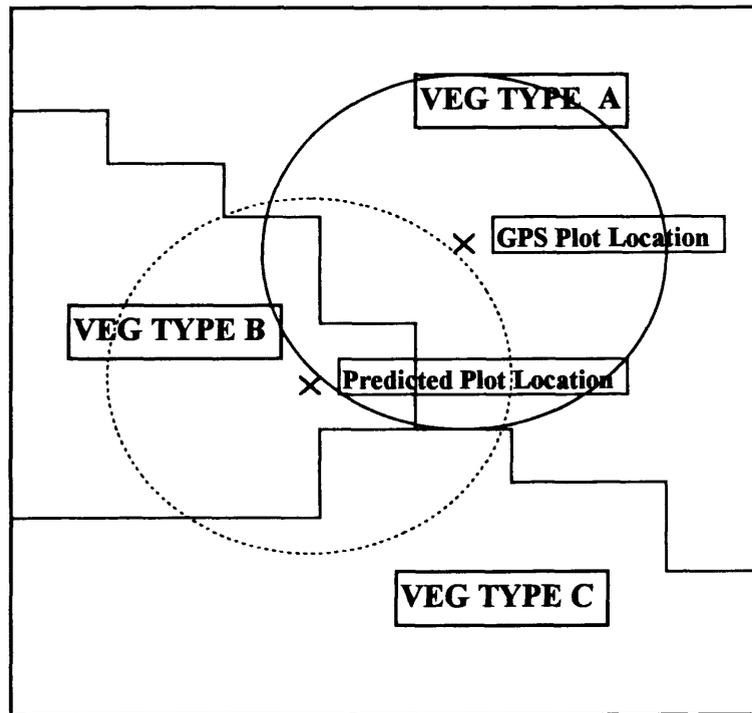


Figure 4. - Example of Predicted vs. GPS plot location.

## ACCURACY ASSESSMENT

### Analysis of Data

Accuracy assessment field data consists of completed forms as shown in Figure 5. Cover type, CALVEG type, size and density are compared to the polygon attributes of the polygon(s) that intersects with the GPS plot location. In the case of multiple-condition plots, the vegetation map is examined for conformity with the condition boundaries indicated on the Plot Diagram (see Figure 3.) An semi-automated approach has been developed for capturing the map label data from the

vegetation map. It involves buffering the GPS plot locations with a given radius, resulting in a polygon coverage of circles roughly one hectare in size. This polygon coverage is overlaid onto the vegetation coverage producing a frequency report listing all the vegetation map labels intersecting each buffered plot. The frequency report is the best indicator of how well the vegetation map compares with the actual vegetation as described by the field crew.

USDA - FOREST SERVICE													REGION 5		
CALVEG ACCURACY ASSESSMENT RECORD															
1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.
X POINT NO.	X CONDITION (A,B,C)	X PRIMARY COVER TYPE	X BEST CALVEG TYPE	X NEXT BEST CALVEG TYPE	X BEST SIZE CLASS	X SECOND BEST SIZE CLASS	X BEST DENSITY CLASS	X SECOND BEST DENSITY CLASS	X SECONDARY COVER TYPE	X BEST CALVEG TYPE	X NEXT BEST CALVEG TYPE	X BEST SIZE CLASS	X SECOND BEST SIZE CLASS	X BEST DENSITY CLASS	X SECOND BEST DENSITY CLASS
1	A	CON	WF	MF	3	2	P	N	SSB	BS	-	-	-	-	-
2	A	CON	WF	MF	3	2	N	P	SSB	BS	-	-	-	-	-
3	A	CON	MF	WF	2	3	N	P	SSB	BS	BL	-	-	-	-
4	A	CON	WF	-	2	-	N	-	SSB	BS	BL	-	-	-	-
5	B	SSB	BS	-	-	-	-	-	CON	WF	-	2	1	S	0

Figure 5. -Example of completed CALVEG data entry form.

## CONCLUSIONS

### Spatial Accuracy

There are several aspects of "accuracy" that can be defined when evaluating forest vegetation maps. Spatial accuracy - or how well a feature on the map matches with a feature on the ground is a key aspect that needs to be evaluated. In forest vegetation, often the best measure of spatial accuracy is how well the boundary between forest and non-forest is represented on the map. A classic example would be a small meadow within a stand of conifers. A perfectly registered vegetation map would depict the meadow in the same location as a GPS traverse of the meadow. Or, as in this case, the condition boundary described by the field crew would match with a polygon boundary in the vegetation map. Because maps derived from TM imagery have errors inherent with registration, as well as other factors, it is very important that the accuracy assessment plot be located with a minimum of error. Using corrected GPS data to properly locate the plots in the GIS is crucial to this effort. The less uncertainty there is in the location of the accuracy assessment plot, the better we can evaluate the accuracy of the map itself.

### **Map Label Accuracy**

The accuracy of map labels in depicting the “right” vegetation type is another aspect to evaluate. This can be done at several levels. The vegetation maps contain a hierarchy of information about vegetation: Life Form/CALVEG/Size & Density. At the most basic level, we need to know how well the map represents Life Forms. Example: how many of the Conifer plots are actually mapped as Conifer?

At the CALVEG level we need to know how many of the Red fir plots are mapped as Red fir? And at the next level we need to know how many of the Large Tree/High Density Red fir plots are mapped as Large Tree/High Density?

The accuracy assessment plots installed on the Modoc National Forest provide the data to answer these questions. They also contain enough “fuzziness” information to allow for interpretation of wrong map labels. If map labels are wrong, we want to know if they are only slightly wrong, or completely wrong. For users of the vegetation map, these issues can be very important.

### **Project Efficiency**

Combining the collection of accuracy assessment data with the standard inventory plot has improved the efficiency of the project. The accuracy data can be collected quickly, once on site. As with most field data collection projects, the largest cost item is getting to the plot. Provided that the plot work can be completed in one day by a two person crew, there is minimal additional expense incurred by collecting the accuracy data. Some additional training is necessary for crews to complete the accuracy assessment form, but the training can be combined with normal inventory procedure training.

Using GPS to precisely locate plot centers is valuable from several standpoints. For the accuracy assessment, it provides control; a position accurate to only a few meters, that can be compared with a map product. If the map contains registration errors, the GPS data will show it conclusively. Secondly, having precise GPS coordinates of the plot location allow for independent evaluation of other map products that describe life form, cover or vegetation. Thirdly, the GPS coordinates provide a navigational aid in relocating the plots for future remeasurement.

### **REFERENCES**

- Forest Inventory and Analysis User’s Guide, 1995. Published by USDA Forest Service, Region 5 - Pacific Southwest.
- Integration of Inventory and Field Data for Automated Fuzzy Accuracy Assessment of Large Scale Remote-Sensing Derived Vegetation Maps in Region 5 National Forests, 1996. Jeff A. Milliken and Curtis E. Woodcock.

# A Method for Measuring the Spatial Accuracy of Coordinates Collected Using the Global Positioning System

Thomas Owens<sup>1</sup> and David McConville<sup>2</sup>

---

**Abstract.**--The study applied a method to measure the spatial accuracy of point data. The quantity of interest was the difference between global positioning system (GPS) coordinates and surveyed coordinates. This difference is a bivariate quantity and the probability distribution function (PDF) can be described by an ellipse with the center at  $\bar{X}$  and  $\bar{Y}$ . The confidence ellipse is an estimate of accuracy; the sample mean is or is not significantly different from the survey locations at a given  $\alpha$ . The tolerance ellipse is an estimate of precision; a given percentage of the population sampled is enclosed in the tolerance ellipse at a given  $\alpha$ . Thirty-five locations were measured with a GPS receiver at surveyed locations in the Mississippi River floodplain near La Crosse, Wisconsin. The average offset was 0.18 m in the easting (X) direction and -1.13 m in the northing (Y) direction. Hotelling's one sample test determined that  $H_0$  (no significant departure from the survey locations exists) was rejected at the 0.05 level, which indicated a systematic error in the sample. Ninety-five percent of the population sampled (at the 0.05 level) was contained in an ellipse centered on (0.18 m, -1.13 m), had a major axis of 7.49 m, a minor axis of 5.12 m, and an angle of 87.74 degrees.

---

## INTRODUCTION

The Environmental Management Technical Center acquires and uses spatial data for inventory and analysis of trends on the Upper Mississippi River in the implementation of the Long Term Resource Monitoring Program. Evaluating accuracy and precision of spatial data in geographic information systems (GIS) is important for determining the appropriate use of these data. However, a good method has not been documented for measuring the accuracy of location measurements because of the problem's complexity and the difficulty in obtaining accurate location information with traditional survey techniques.

---

<sup>1</sup> Cartographer, Environmental Management Technical Center, Onalaska, WI.

<sup>2</sup> Professor, St. Mary's University, Winona, MN

It is conceptually easier to determine the spatial accuracy of points rather than lines or polygons. A point may represent a specific location such as a survey point or it may represent an area displayed at a small scale. If a point represents a real single location, it is possible to measure its displacement from its "true" location. Lines and polygons are complex abstract objects and measuring their spatial accuracy is more complex (Goodchild and Gopal 1989). For this reason, this study concentrates on the spatial accuracy of points and leaves the spatial accuracy of polygons for later study.

The advent of the Global Positioning System (GPS) has reduced the expense and complexity of traditional survey techniques for measuring the location of objects on the earth. The GPS enables nonsurveyors to determine the location of objects with relative ease, but gives no information about the data's accuracy and precision. This study devises a straight-forward, repeatable, and statistically sound method of estimating the horizontal accuracy of GPS derived coordinates. Two measures of accuracy are widely used, RMS and CEP. Root Mean Square (RMS) error is the square root of the sum of the square of the euclidian distances of the error, divided by the number of observations. Circle Error Probable (CEP), is defined as a circle that will enclose 50 percent of the dot print. These measures have two disadvantages: (1) they contain no information about the direction of the error, and (2) they are only measures of accuracy, with no information about the variability (precision) of the measurements.

Error and accuracy are conceptually the opposite sides of the same coin. Error is the difference between a quantity's correct value and its estimated value (Thapa and Bossler 1992). Accuracy is the nearness of quantities to their true values (Bolstad et al. 1990). Errors may be classified into three types: (1) gross errors or blunders, (2) systematic errors, and (3) random errors. Gross errors are caused by carelessness or mistakes of the observer using the equipment. Systematic errors are due to environmental factors, instrumental imperfections, and human limitations; they introduce bias into the estimate. Even after all gross and systematic errors have been eliminated, some small random errors will remain.

Precision is the similarity of repeated measurements among themselves. In a sample of coordinates representing the same position, precision is the deviation of the points from the mean coordinate. Total error for a particular measurement is the sum of the mean (systematic) and deviation (random) errors. Average error (accuracy) and distribution about this mean (precision) characterize the positional uncertainty of digital spatial data, which is related to the statistical distribution of the errors and can be modeled by a probability distribution function (PDF). The average difference between the test and control data estimates the accuracy, while the frequency distribution establishes the theoretical PDF which contains information on both accuracy and precision.

When a sample of GPS coordinates is taken at surveyed locations, the quantity of interest is the difference in GPS coordinates from the surveyed coordinates. The X and Y difference between GPS coordinate estimates and the surveyed coordinate is a bivariate quantity. The joint PDF can be described by a probability ellipse (Batschelet 1981) with the center at  $\bar{X}$  and  $\bar{Y}$ . An ellipse is an appropriate shape for a joint PDF because it has two dimensions; it is not rectangular because the joint probability of points occurring in the corners is small, and it is generally not circular, because X and Y are not necessarily the same. The ellipse can be described with five statistics: the sample means of (1) X, (2) and Y, the sample standard deviations of (3) X, (4) and Y, (5) and the correlation coefficient between X and Y. The quantities X and Y are jointly distributed, where X and Y depend on each other, but different pairs are independent of each other in a sample.

There are three ellipses of interest: the standard ellipse, the confidence ellipse, and the tolerance ellipse. The standard ellipse is a descriptive tool used to visualize the shape of the ellipse and its orientation. It is readily calculated and contains about 40% of the sample population. It is not dependent on the sample size and cannot be used for statistical inference. Five statistics are required to construct the ellipse:  $\bar{X}$ ,  $\bar{Y}$ ,  $s_x^2$ ,  $s_y^2$ , and the correlation coefficient  $r$  ( $r = \text{Cov}(X,Y)/s_x s_y$  and  $\text{Cov}(X,Y) = 1/(n-1)\sum(X_i - \bar{X})(Y_i - \bar{Y})$ ), which measures the joint behaviors of X and Y. The center of the ellipse is  $(\bar{X}, \bar{Y})$  and the equation of the ellipse is as follows:

$$s_y^2(X - \bar{X})^2 - 2rs_x s_y(X - \bar{X})(Y - \bar{Y}) + s_x^2(Y - \bar{Y})^2 = (1 - r^2)s_x^2 s_y^2 \quad (1)$$

For ease of calculation, some coefficients are shortened:

$$\begin{aligned} A &= s_y^2 & B &= -rs_x s_y & (2, 3) \\ C &= s_x^2 & D_s &= (1 - r^2)s_x^2 s_y^2 & (4, 5) \end{aligned}$$

An ellipse is defined by a major axis (a), a minor axis (b), and an angle  $\theta$  of translation, which is the angle that the major axis is offset from the X axis (and the angle the minor axis is offset from the Y axis).

Another coefficient must be calculated to arrive at the ellipse parameters:

$$R = [(A - C)^2 + 4B^2]^{1/2} \quad (6)$$

Then we can calculate the ellipse parameters:

$$a = [2D/(A + C - R)]^{1/2} \quad (7)$$

$$b = [2D/(A + C + R)]^{1/2} \quad (8)$$

$$\theta = \arctan[2B/(A - C - R)] \quad (9)$$

The other two ellipses have an identical shape and orientation but have different major and minor axes. The Hotelling's confidence ellipse covers the sample's center with a given accuracy and estimates accuracy. In addition, Hotelling's one-sample test can be used to determine if the bivariate sample mean is significantly different than 0. The confidence ellipse is a region that contains the population mean at a given probability with the center of the ellipse at  $(\bar{X}, \bar{Y})$ . The equation of the ellipse is as follows:

$$s_y^2(X - \bar{X})^2 - 2rs_x s_y(X - \bar{X})(Y - \bar{Y}) + s_x^2(Y - \bar{Y})^2 = (1 - r^2)s_y^2 s_x^2 n^{-1} T^2 \quad (10)$$

Note that the right side of the expression has an added term,  $T^2$ , which is based on the familiar F value for the univariate solution where:

$$T^2 = 2[(-1)/(n - 2)]F_{2,n-2} \quad (11)$$

$F_{\alpha/1, n-2}$  denotes the critical value from the F one-tailed distribution with n-2 degrees of freedom and significance level of  $\alpha$ . The confidence ellipse is then calculated using the same coefficients described for the standard ellipse.

The tolerance ellipse is an estimate of precision. It is an estimate of the confidence (e.g., 95%) of a percentage of the population sampled (e.g, 95%) that is enclosed in the tolerance ellipse. The tolerance ellipse is a region that contains a given percent of the population at a given probability with the center of the ellipse at  $(\bar{X}, \bar{Y})$ . The equation of the ellipse is:

$$s_y^2(X - \bar{X})^2 - 2rs_x s_y(X - \bar{X})(Y - \bar{Y}) + s_x^2(Y - \bar{Y})^2 = (1 - r^2)s_y^2 s_x^2 H \quad (12)$$

The right side of the expression has the term H, which is based on the non-central  $\chi^2$  distribution and is approximated by Table 2 in Chew (1966). The tolerance ellipse is then calculated using the same coefficients described for the standard ellipse.

This elliptical method of describing spatial accuracy and precision is valid because it is (1) intuitive, (2) understandable graphically, (3) statistically sound, (4) sufficient (e.g., it describes all relevant parameters), and (5) concise.

## METHODS

The location of the study included Navigation Pools 7 and 8 of the Upper Mississippi River near La Crosse, Wisconsin. This reach of the river is a complex of islands, channels, and backwaters surrounded by 500-ft high bluffs.

Forty-one surveyed ground locations were used as test points. Thirty-six third-order points were surveyed by the U.S. Army Corps of Engineers and were distributed along the main channel of the river. Five first-order points located at road intersections in the floodplain were surveyed by the Wisconsin and Minnesota Department of Highways.

A Trimble Pathfinder Basis Plus GPS receiver was used in the field as the rover. The GPS measurements were taken with the rover on June 8, 9, and 17, 1994. The PDOP mask was set to 5. The receiver was set to take one position per second and the horizon mask was set at 10°. The unit received signals until approximately 200 readings were taken at each point. A Trimble Pathfinder Professional receiver was established at a surveyed point to act as the base station. The base was set to take positions once every 5 seconds, had a PDOP mask of 5, and a horizon mask of 10°. The measurements were post-processed to introduce differential correction using Trimble's PFinder software on a personal computer.

Survey measurements were recorded in the State Plane Coordinate System, Minnesota South zone, and converted to the Universal Transverse Mercator (UTM) projection, zone 15. The GPS measurements were taken in WGS-84 and converted to the UTM projection, zone 15, NAD27.

The X (easting) and Y (northing) differences between the survey GPS measurements were tested to determine if their distribution was normal using the  $\chi^2$  (chi-square) test (Zar 1984). If the easting and northing differences were normally distributed, it is possible to use bivariate normal tests to test for systematic bias from zero and to describe the PDF with ellipses (Batschelet 1981). Hotelling's confidence ellipse, the standard ellipse, and the tolerance ellipse (Chew 1966) were calculated to describe the PDF. Hotelling's one-sample test was run to determine if the sample average was significantly different than 0.

## RESULTS AND DISCUSSION

Thirty-five usable locations were differentially corrected from the 41 locations measured. Five of the original locations were not usable because the base station and rover did not receive positions from the same satellites and so the GPS readings were not corrected during differential post-processing (possibly due to setting the mask angle to 10° on both the rover and the base station). One surveyed point had incorrect coordinates and was discarded.

The average distance from the survey points was -1.13 m in the northing (Y) direction and 0.15 m in the easting (X) direction.

The easting and northing distances were tested for normality.  $H_0$  (that the distributions were normal) was not rejected at the 0.05 level. Thus, we could calculate the descriptive and statistical ellipses and test to determine if there was a significant difference from 0 for the mean distance.

The standard ellipse was centered on the sample means of (0.15, -1.13), and had a major axis of 2.60 m and a minor axis of 1.77 m with an angle from the X axis of  $87.74^\circ$  (figure 1). The confidence ellipse had a major axis of 1.12 m and a minor axis of 0.77 m, with the same angle of  $87.74^\circ$ . Hotelling's one-sample test was used to determine if the sample means were significantly different than 0. The result was that  $H_0$  (that the population means equal 0) was rejected at the 0.05 level. The tolerance ellipse, which was calculated to contain 95% of the population at the 95% confidence level, had a major axis of 7.49 m and a minor axis of 5.12 m.

The confidence ellipse does not encompass the origin (0, 0), which graphically shows that the sample mean is different than 0. Thus, the sample has systematic error that is offset in the south and east directions. We can also say that we are 95% confident that 95% of the population we sampled is contained in an ellipse that is centered on 0.15, -1.13, and has a major axis of 7.49 m, a minor axis of 5.12 m with an angle of  $87.74^\circ$  at the 0.05 level.

## CONCLUSION

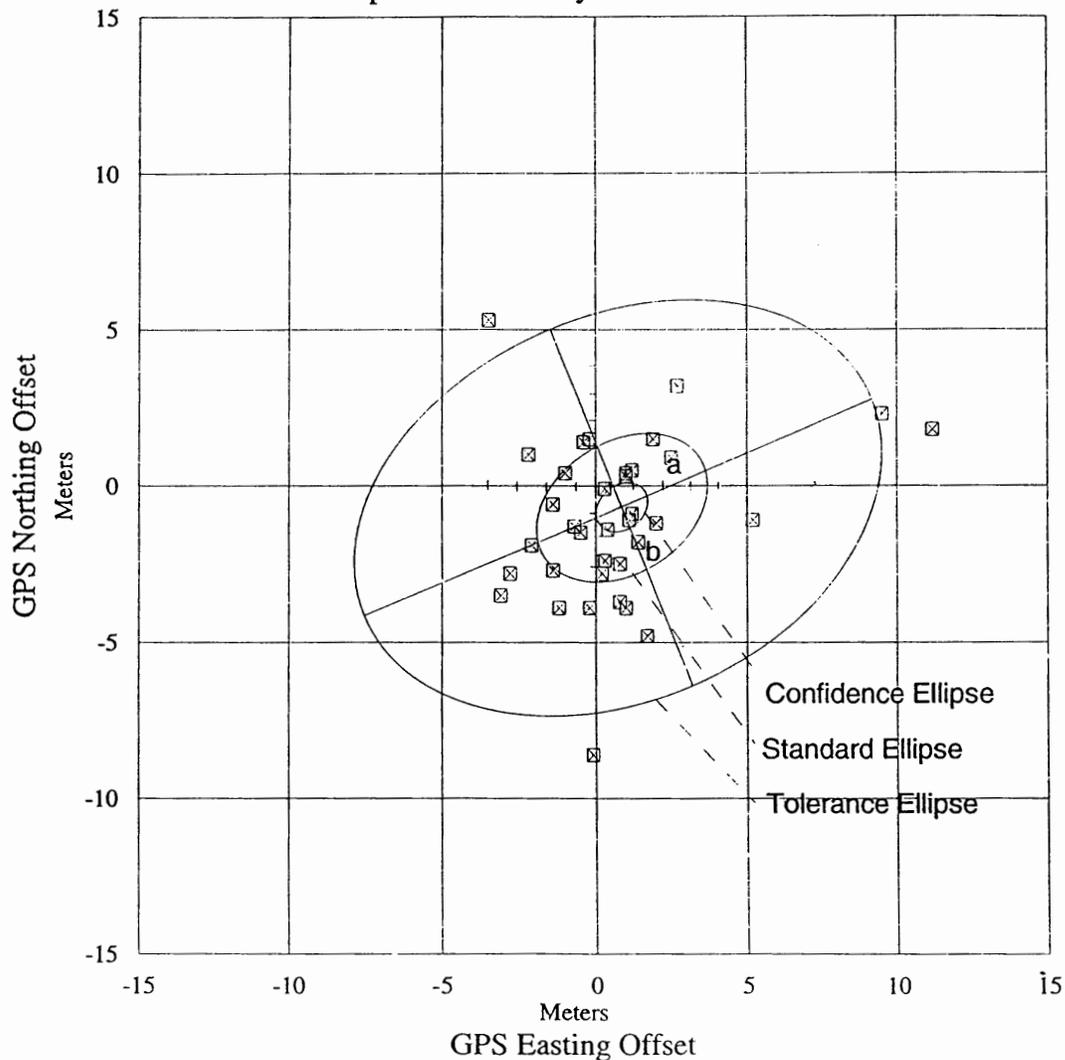
The purpose of this study was to apply a method to measure the spatial accuracy of point data. Hotelling's one-sample test made inferences about the sample data's means and a tolerance ellipse described the PDF of the population. The confidence ellipse is an intuitive graphic that shows whether systemic error exists in the sample. This method is theoretically sound and provides intuitive and concise statistics about a population. We can now describe the accuracy of the sample data generated by GPS described in the study as having a mean that is 0.18, -1.13 m, and has systemic error in the east and south directions. We can further state that 95% of the population is contained in an ellipse (at the 95% confidence level) with a center at 0.18, -1.13 m having a major axis of 7.49 m, a minor axis of 5.12 m, and an angle of translation from the horizontal of  $87.74^\circ$ .

This method to measure and describe the spatial accuracy of point data is useful for three reasons. First, the method is repeatable; similar results will be obtained for a given set of points when the method is repeated and accuracy statements can be compared for different sets of data. Second, it is statistically valid; this method provides a probability statement about the accuracy statement. Third, it is sufficient. Merely stating that the GPS-derived points are "within 5 m of their

true location" is not sufficient because no probability statement is associated with the accuracy statement. In addition, X and Y coordinates are related, but do not vary in a perfect one-to-one relationship; it is necessary to describe how both X and Y vary, along with the confidence ellipse, which describes the accuracy, and the tolerance ellipse, which describes the precision.

## GPS Offsets from Survey Points

Comparison of Survey Points and GPS Points



**Figure 1.**--Distribution of distances between the GPS-derived coordinates and survey-derived coordinates and ellipses which describe the sample, accuracy, and probability distribution function.

## ACKNOWLEDGEMENTS

This study was supported by the U.S. Army Corps of Engineers and the National Biological Service through the Upper Mississippi River System Long Term Resources Monitoring Program.

## REFERENCES

- Batschelet, E. 1981. Circular statistics in biology. Academic Press, New York. 310 pp.
- Bolstad, P., P Gessler, and T. Lillesand. 1990. Positional uncertainty in manually digitized map data. *International Journal of GIS*. 4(4): 399-412.
- Chew, V. 1966. Confidence, predictive, and tolerance regions for the multivariate normal distribution. *American Statistical Association Journal* 61:605-617.
- Goodchild, M., and S. Gopal. 1989. Preface. Pages xi-xv *in* National Center for Geographic Information and Analysis. M. Goodchild and S. Gopal, editors. Accuracy of Spatial Database. Taylor and Francis, New York.
- Thapa, K., and J. Bossler. 1992. Accuracy of spatial data used in GIS. *Photogrammetric Engineering and Remote Sensing* 58(6):835-841.
- Zar, J. 1984. Biostatistical analysis. Prentice-Hall, Englewood Cliffs, New Jersey. 620 pp.

## BIOGRAPHICAL SKETCHES

Thomas Owens is a Cartographer with the National Biological Service's Environmental Management Technical Center in Onalaska, Wisconsin. He received a Master of Science in forest management and remote sensing from the University of Minnesota in 1978. He is Leader of the Cartographic Services Group and provides GIS expertise to the EMTC.

David McConville is a codirector of the Resource Studies Center and Professor of Biology at Saint Mary's University of Minnesota. He holds a Ph. D. in fisheries from the University of Minnesota and has over 25 years of experience working on aquatic ecosystems and, in particular, Mississippi River ecology.

# GPS vs Traditional Methods of Data Accuracy Input: Improving Spatial Data Accuracy?

Russell Combs, Jr.<sup>1</sup>, James L. Smith<sup>2</sup>, and Paul V. Bolstad<sup>3</sup>

---

**Abstract.**—Geographic information systems (GIS) have become an integral part of many natural resource organizations. The increased interest and use of these systems for a variety of managerial tasks has resulted in an increased concern about the accuracy of spatial databases. Many natural resource organizations are using their GIS to perform complex analysis and they are becoming very aware of what the cost might be for using inaccurate spatial data. Thus many organizations are involved in developing procedures or methods for improving and maintaining the quality and accuracy of their spatial data. This paper will focus on the issue of spatial data accuracy and how it can be improved or maintained by natural resource organizations. In particular, the paper will concentrate on one of the main areas influencing spatial data accuracy, data input. The paper will review and compare the positional accuracy levels of some of the most common methods of data input (manual digitizing) to newer technology and techniques (GPS, PC-based single-photo space resection) now available to natural resource organizations.

---

## INTRODUCTION

Positional accuracy in a natural resource GIS can be defined as a measure, usually in ground distance, of how well the digital object coordinates in the various spatial data layers correspond to the “true” coordinates of an entity on the ground (Bolstad & Smith 1992, Chrisman 1991, and Blakemore, 1984). The accuracy of a spatial database is a direct function of the methods used to input the spatial data, and the sources of data used to collect the spatial information. As with any mapping procedure, whether digital or cartographic, generalizations are made and database coordinates used to represent entities may differ from their “true” ground locations. Thus, the accuracy of a spatial database is inherently limited by the method or sources of data used to construct the database.

A natural resource GIS spatial database usually includes multiple data layers, with each layer representing a different theme or group of similar features.

1 2 3

---

<sup>1</sup> Database/GPS Forester, Canal Forest Resources, Inc., Charlotte, NC

<sup>2</sup> Manager GIS and Remote Sensing Services, Canal Forest Resources, Inc., Charlotte, NC

<sup>3</sup> Assistant Professor, Department of Forest Resources, University of Minnesota

The individual data layers in a natural resource GIS are not typically developed or created using the same methods or sources of spatial data. Natural resource organizations routinely deal with ground features that are not well mapped or must work with with poorly or undocumented maps. Thus, a wide range of data sources are often used, and often two different sources of data are used within a given data layer. This leads to a combination of methods and sources being used to develop or update the spatial database, and in turn leads to a spatial database with many different data layers having varying levels of positional accuracy. This paper will review the accuracies of the most common data input methods/sources, and propose some suggestions on how natural resource organizations can improve the positional accuracy of their spatial databases. In addition, the paper will discuss what if, any accuracy or benefit may be gained by using some of the new technology becoming available.

### **Traditional Methods/Sources of Spatial Data**

The traditional methods and sources of spatial data input for natural resource organizations can be categorized into two main types: 1) manually digitizing from large-scale maps or aerial photographs, and 2) the use of purchased digital data (DLG, TIGER, DEM). While purchased digital data is becoming more widespread throughout the natural resource community, its use is limited to those organizations with established GIS systems, and those who have familiarity with these type of data. Manual digitization is still by far the most common method of data input, and thus the major source of positional error in spatial databases. Positional errors resultant from manual digitization are generally a function of three sources: media limitation, user error, and control point or coordinate registration error (Walsh *et al*, 1987).

While all three of these factors vary as to the affect they have on the positional accuracy of manually digitized data, the first of these factors, source media limitation, has the largest impact on the amount of positional error introduced to spatial data during manual digitization. Manually digitized data in natural resource organizations are usually collected from either aerial photography or large-scale maps. In both cases the source data has inherent limitations that affect the accuracy of the data. Aerial photography is affected by tilt and terrain displacement, while many large-scale natural resource maps contain significant positional errors due to poor drafting, incorrect boundary placement, or drafting from aerial photography.

The amount of positional error that can be introduced to spatial data because of these factors has been documented, and by reviewing previous research, an estimate of the positional error caused by manual digitization can be obtained. For example, horizontal positional errors of up to 73.0 meters in steep terrain and

15.0 meters in low terrain have been observed for spatial data digitized directly off of aerial photography (Bolstad, 1992). On the other hand, large scale maps often have linear or areal features that have line widths which vary between 0.25 - 1.01 mm (Bolstad & Smith, 1992), corresponding to widths on the ground of anywhere between 6.0 - 24.0 meters if the data were collected from a 1:24,000 scale map. Another study observed line locations to vary by 0.127 mm of their true positions on a 1:24,000 scale map, equivalent to 3.04 meters in actual ground distance (Dunn *et al*, 1990). As can be seen, it does not take much variation in a source document to start adding up to large amounts of inaccuracy on the ground. A factor that is sometimes overlooked when dealing with linear and positional accuracies, is the associated area accuracy that is directly related to the linear accuracy of the spatial database. For instance Chen and Finn (1994) stated that average manually digitized area error is low, with manually digitized data generally underestimating polygon area by 1-3%. In addition, Wiles (1988) observed acreage errors of up to 10.23% for data digitized directly from aerial photography in varying terrain.

The second source of positional error resultant from manually digitization is from operator or user error. Operator error is the error resultant from the operator's inability to trace linear or point features with the digitizing puck. Operator digitizing ground errors of up to 1.15 m in the x-direction and 1.17 m in the y-direction have been documented for manually digitized data from 1:24,000 scale maps (Combs 1995, Warner & Carson, 1991). In addition, operator inaccuracies of 0.95 m in the x-direction and 1.26 m in the y-direction have been calculated for manually digitized data collected from NAPP photography (Combs, 1995).

In addition to source errors and user errors, positional accuracy during manual digitizing is also affected by control point or coordinate registration. This is an issue that is often overlooked by natural resource organizations, but poor control point registration does affect the overall accuracy of the digitized data. A few studies have highlighted the importance of using accurate control points. Fernandez *et al*. (1991) documented map derived control point errors ranging from 2.13 to 7.62 meters in ground distance. Bolstad *et al* (1990) observed positional error attributed to poor control points ranging from 1.52 to 85.04 meters on the ground for data digitized directly from aerial photography in varying terrain. Control point registration and accuracy can be increased by using well known established point locations that have been surveyed or through the use of global positioning systems (GPS). Thus, if the positional errors associated with manual digitization are additive, which they most likely are to some extent, spatial data collected using these methods could be tens to hundreds of meters off in positional accuracy.

## Non-traditional Methods/Sources of Spatial Data

While manual digitization and the use of purchased digital data are still popular methods and sources of spatial data, the development of new data input methods and techniques may provide natural resource organizations with more accurate and cost-effective alternatives. Two of these new techniques or methods will be discussed: 1) GPS and 2) PC-based single-photo space resection. While both of these alternatives are not necessarily "new", they have not been extensively applied in natural resource organizations.

GPS is a satellite-based positioning system which operates using L-band radio signals to provide highly accurate position, velocity, and time data. GPS can provide accuracy levels of up to 30.0 cm (P-code) and 1-5 meters (Course/Acquisition Code or C/A code). Civilian users can achieve accuracies equal to those of the P-code but generally this accuracy is reserved for military use or surveyors. Single-photo space resection is accomplished by measuring photo-coordinates, via a digitizing tablet, and applying the collinearity equations to these data to calculate ground coordinates. Since the procedure is based upon the photogrammetric principles of space resection, the calculated ground coordinate accuracy is improved due to the reduction of the tilt and terrain effects inherent in the photograph.

Recently studies have been conducted to determine the accuracy of GPS and single-photo space resection in a natural resource setting. Evans *et al.* (1992) reported average positional accuracy of 2.01 m between GPS positions and actual forest plot center positions during navigation trails to locate known forest plot centers. In other studies, applying differential correction to the mean of 100 to 300 position fixes resulted in positional ground accuracies of between 2.0 to 3.0 m (August *et al.*, 1994) and 2.0 to 4.0 m (Deckert, 1994). A previous study reported positional ground errors from single-photo space resections of two different mountainous study areas to be 11.39 m and 7.07 m respectively (Ran, 1992). That would translate to about 60.0 m more accurate than the data collected directly from aerial photography in steep terrain as reported in the studies previously discussed.

A more recent study involved comparing the accuracies of point features manually digitized from topographic maps and NAPP photography versus the accuracies of the same features collected using single-photo space resection and GPS techniques (Combs, 1995). For this study, C/A code level GPS was used to determine the "true" ground or reference location of the test features. While the GPS data collected for this study can't determine the actual ground locations of the test features, previous work has determined C/A code level GPS points to be within 1.0 - 3.0 meters of the "true" ground location (Deckert, 1994). For each

test point, 300 position fixes were collected and averaged after differential correction to determine a mean positional location for the test point. The manually digitized and single-photo space resection data were then compared to the GPS value and the difference was determined to be the amount of positional error attributed to that method/source. Table 1 highlights the results of this study and shows the average positional accuracy that was determined by using the various techniques and sources.

**Table 1. Mean positional error (meters) for two study sites in Virginia, GPS used as "true" ground location.**

Data Set	High-Relief Study Site	Low-Relief Study Site
Space Resection	6.27	5.64
Digitized Aerial Photograph	25.86	7.14
Topographic Map	11.20	11.64

The results in Table 1 show how spatial accuracy can be improved by using either GPS or the space resection technique. The positional accuracy of data digitized directly from aerial photography is about what would be expected. As shown in Table 1, the high-relief study site had a mean positional accuracy of more than 3 times greater than the low-relief study site. This difference is attributed to the greater amount of terrain relief and tilt displacement that is present in the high-relief study site aerial photography. In contrast, the single-photo space resection data sets provided mean positional accuracies that are almost equal between the two study sites. Since the space resection reduced the tilt and terrain effects inherent in the aerial photography (and assuming all other factors are constant), the resultant mean positional accuracies were similar and improved over just digitizing directly from the aerial photography.

The digitized topographic map data from both study sites provided similar positional accuracies that were greater than the NMAS (National Map Accuracy Standards), but both were still more than half of what was obtained by the space resection technique. Other results that are not directly shown in Table 1 are the accuracy of the GPS data. While GPS was only used for the "true" ground reference value for this study, the results in Table 1 show how much each of the other data collection techniques differed in positional accuracy from the GPS values. For example, on the high-relief study site the GPS data was on average 25.0 meters more accurate than digitizing directly from aerial photography. In addition, the GPS data proved to be on average 11.0 meters more accurate than digitizing directly from 1:24,000 scale topographic maps. While the GPS proved to provide the greatest positional accuracy, both it and the single-photo space resection increased the reliability and confidence of the data collected in this study.

## CONCLUSIONS

The accuracy of digital spatial data collected by natural resource organizations is affected by the source of the data and the method use to collect the data. As the use of GIS continues to grow in natural resource organizations, the reliability and accuracy of the data used in GIS analysis is becoming increasingly important. Greater than any other time in recent history, natural resource organizations are being held responsible for their actions and the decisions they make concerning the environment. The research and results reviewed in this paper show that spatial data collected using the traditional methods may adversely affect the accuracy of spatial databases. Research presented in this paper showed that data digitized directly from large-scale maps and aerial photography can provide positional accuracies anywhere between 6.0 to 73.0 meters. On the other hand, spatial data collected using a PC-based single-photo space resection technique can provide positional accuracies between 5.0 to 7.0 meters, depending on the amount of terrain relief. In addition, GPS collected spatial data has been shown to be capable of providing positional accuracies from 1 to 4 times more accurate than some of the traditional techniques discussed in this paper. As the use of GIS technology continues to expand in natural resource organizations, a PC-based single-photo resection or more particular the use of GPS, offers an alternative to traditional methods of data entry and can provide reliable and more accurate spatial data.

## REFERENCES

- August, P. , J. Michaud, C. Labash, and C. Smith. 1994. GPS for Environmental Applications: Accuracy and Precision of Locational Data. *Photogrammetric Engineering & Remote Sensing*. 60(1):41-45.
- Blakemore, M. 1984. Generalization of Error in Spatial Databases. *Cartographica*. 21:131-139.
- Bolstad, P.V. 1992. Geometric Errors in Natural Resource GIS Data: Tilt and Terrain Effects in Aerial Photographs. *Forest Science*. 38(2):367-380.
- Bolstad, P.V., P. Gessler, and T.M. Lillesand. 1990. A Variance Components Analysis of Manually Digitized Map Data. *Proc. ACSM-ASPRS*. Vol 3. 9-18pp.
- Bolstad, P.V., and J.L. Smith. 1992. Errors in GIS: Assessing Spatial Data Accuracy. *Journal of Forestry*. 90(11):21-29.
- Chen, Z. Z., and J.T. Finn. 1994. The Estimation of Digitizing Error and Its Propagation with Possible Application to Habitat Mapping. *Proc. International Symposium on the Spatial Accuracy of Natural Resource Databases*. 57-66 pp.
- Chrisman, N.R. 1991. The Error Component of Spatial Data. *Geographical Information Systems, Volume 1: Principles*. John Wiley & Sons. 165-173 pp.
- Combs, R.G., 1995. Positional Accuracy in a Natural Resource Database: Comparison of a Single-Photo Resection Versus Affine Registration. Master's Thesis. VPI&SU. 148 pp.

- Deckert, C.J. 1994. Canopy, Terrain, and Distance Effects on Global Positioning System Position Accuracy. Master's Thesis, VPI&SU. 70 p.
- Dunn, R., A. R. Harrison, and J.C. White. 1990. Positional Accuracy and Measurement Error in Digital Databases of Land Use: An Empirical Study. *International Journal of Geographical Information Systems*. 4(4):385-398.
- Evans, D.L., R.W. Carraway, and G.T. Simmons. 1992. Use of Global Positioning System (GPS) for Forest-Plot Location. *Southern Journal of Applied Forestry*. 16(2):67-70.
- Fernandez, N., D. F. Lozano-Garcia, G. Deeds, and C. J. Johannsen. 1991. Accuracy Assessment of Map Coordinate Retrieval. *Photogrammetric Engineering & Remote Sensing* 57(11):1447-1452.
- Ran, L., 1992. Single Digital-Photo Correction for a GIS Application and Error Analysis. Master's Thesis, VPI&SU. 139 pp.
- Walsh, S. J., D. Lightfoot, D. Butler. 1987. Recognition and Assessment of Error in Geographic Information Systems. *Photogrammetric Engineering and Remote Sensing*. 53(10):1423-1430.
- Warner, W. S., and W. Carson. 1991. Errors Associated with a Standard Digitizing Tablet. *ITC Journal*. No. 2:82-85 pp.
- Wiles, S. J. 1988. Evaluation of Photographic Properties for Area Estimation. Master's Thesis. VPI&SU. 100 p.

### **BIOGRAPHICAL SKETCH**

Russell G. Combs, Jr. is a Database/GPS Forester with Canal Forest Resources, Inc. in Charlotte, NC. Russ holds a B.S. in forestry and a M.S. in GIS and GPS from Virginia Tech. Russ has been employed with CFR for the past year where he is responsible for managing the inventory database for CFR's clients as well as heading up Canal's company wide GPS program and assisting on various GIS projects and analysis.

James L. Smith holds a B.S. in Forestry and an M.S. in Forest Biometrics from the University of Georgia, and a Ph. D. in Forest Biometrics and Remote Sensing from Virginia Tech. He was a member of the faculty of the Virginia Tech Department of Forestry at Virginia Tech for 13 years, where he specialized in quantitative GIS and remote sensing issues. For the last two years, Jim has been the Manager of the GIS and Remote Sensing Group at Canal Forest Resources, Inc. in Charlotte, NC.

Paul V. Bolstad is an Assistant Professor in the Department of Forest Resources at the University of Minnesota. Paul holds a B.S. in Forestry from the University of California at Berkeley, a M.S. in Forestry from N.C. State, and a Ph. D. in Forestry from Wisconsin University. Paul was a member of the faculty of the Virginia Tech Department of Forestry at Virginia Tech for 6 years, where he specialized in GIS, Remote Sensing, and GPS issues.

# Large Scale Tropical Forest Change Monitoring Using Multiple Resolution Satellite Data : from Hot Spot Detection to Global Deforestation Assessment ?

Hervé Jeanjean<sup>1</sup>, Frédéric Achard<sup>2</sup> and Jean-Paul Malingreau<sup>3</sup>

---

**Abstract** - The rate of tropical forest degradation and deforestation is a serious threat on environmental issues, socio-economic balances and land use planning and is considered as a major concern for decision-makers at global, regional and local scales. Appropriate tools are needed to address this multiscale and multipurpose issue. Sampling techniques can be applied in large forest surveys using high resolution satellite data. "Wall to wall" coverage is a feasible alternative, but due to the high cost of data, low resolution satellite data are more adapted to this method, provided that estimations are corrected using multistage procedures. Recent advances have proved that this approach is feasible for mapping and assessing tropical forest on large scale with reasonable accuracy. However, for forest change detection and monitoring, few methods have been developed so far enabling the estimate of change at global scale. In this paper, change detection based on a combination of LRSD such as AVHRR data and HRSD such as TM and SPOT data is addressed. A preliminary investigation is carried out in a pilot area in Southeast Asia where changes detected with HRSD (hot spot areas) are compared with AVHRR forest class proportions and fragmentation patterns. A discussion is then proposed on the necessary steps towards an operational system for global tropical forest monitoring.

---

## INTRODUCTION

Tropical forests constitute fragile ecosystems which shelter most of the world animal and vegetal species. Playing an essential role in global environmental and climatic balance, tropical forests are threatened by a fast disappearance with dramatic consequences on biodiversity loss, shortage of wood and non-wood products, soil degradation and erosion, deterioration of hydrological systems, disruption of local socio-economic development...

According to the last estimates of FAO FRA 1990, 15.4 millions ha of tropical forests have disappeared each year between 1980 and 1990, which represent an average deforestation rate of 0.8 % of total tropical forest resources. In 1980, tropical forests were estimated at 1,910 millions ha, and 1,756 millions ha in 1990. The deforestation mechanism has become a critical issue not only for the countries where this problem is a real concern, but also for the international scientific community who has launched

---

<sup>1</sup> Remote sensing and GIS expert in forestry, SCOT Conseil, 1 rue Hermès, F-31526 Ramonville France

<sup>2</sup> MTV Group, TREES Project Manager, Institute of Remote Sensing Applications, I-21020 Ispra Italy

<sup>3</sup> Head of Monitoring Tropical Vegetation Unit, TREES Project Coordinator, Institute of Remote Sensing Applications, I-21020 Ispra Italy

important research programmes to better evaluate its exact magnitude and to improve the knowledge of the different factors interacting in this complex processus. Intensive logging practices, shifting cultivations and conversion of forest lands to agricultural areas are some of the main causes, direct or indirect, of deforestation. Above all, the modes of appropriation of forest lands by local farmers represent in many cases a critical issue which interacts with legal aspects of land management. To tackle the environmental effects of deforestation, one of the necessary actions consists in building up a continuous monitoring systems along three major steps : (1) development of sound methods for estimating forest areas (static approach), (2) development of sound methods for estimating forest cover changes on a quantitative and structural basis (dynamic approach) and (3) modeling of deforestation processes (anticipating approach). Considerable advances have been made on the development of methods for retrieving forest proportions based on CRSD. Few methods have been developed for change detection and quantification, and they are generally using sampling techniques with HRSD. The modeling of deforestation requires to consider many parameters which encompass not only the vegetation characteristics of the area, but also physical and anthropic characteristics.

This paper is mainly addressing the second issue, after reviewing some techniques for global estimates of forest areas.

## **GLOBAL ESTIMATE OF FOREST COVER PROPORTION**

### **The main actors**

As far as global assessment is concerned, CRSD such as AVHRR data can provide valuable information for forest resources assessment. Future sensors (vegetation instrument on board SPOT 4 and MODIS) are expected to give further possibilities in this field.

Several institutions or programmes have already started significant activities in global survey of tropical forests : (i) the European Union with the TREES Project (TRopical Ecosystem Environment observation by Satellites) where wall to wall coverage is made using AVHRR data with a sample of HRSD for calibration purposes, (ii) FAO with the Forest Resources Assessment 1990 project where TM data are sampled over the entire tropical belt, (iii) IUFRO (International Union of Forest Research Organizations) which has published some guidelines for forest monitoring (Päivinen, 1994), (iv) NASA which is carrying out several projects in South America and Africa (Landsat Pathfinder project in the framework of the Global Change Research Programme : Chomentowski et al., 1994), (v) UNEP/GRID which has undertaken some global analysis with AVHRR data in the amazon basin, and has been involved with TREES in Southeast Asia, (vi) the Woods Hole Research Centre which has realized a map of South America using different sensors, (vii) IUCN (International Union for Conservation of Nature and Natural Resources) and WCMW (World Conservation and Monitoring Centre) which published in the 1990's the Conservation Atlas of Tropical Forests, (viii) IGBP (International Global Biosphere Programme) which is working on global change issues. Over Africa, other regional initiatives with environmental components are in progress or in preparation : the CARPE (Central African Regional Program for Environment), and RIMP (Regional Information Management Programme).

## **Existing methods for global forest survey**

Four different methods can be applied for estimating tropical forest areas on a global basis (Czaplewski, 1992). A first approach is consisting in compiling forest statistics provided by national or regional forest inventories. FAO used this approach in 1980 and to a limited extent in 1990. The main problems of this method are the discrepancies between the nomenclatures, the variables, the accuracy in the sampling designs, the dates of the inventories. A second method is based on the analysis of a wall to wall coverage of CRSD data (AVHRR data). In this approach, forest resources can be mapped, e.g. located, and estimated using some AVHRR classification schemes. However, the estimation of forest proportion is presenting a systematic bias due to aggregation effects. The magnitude of this error depends on three factors : the spatial resolution of the map, the initial proportions of forest in the landscape, and the spatial arrangement of forest at fine resolution (Turner et al., 1989; Moody and Woodcock, 1994). It is therefore recommended to apply a correction function which may be derived from a double sampling design with regression estimator. The regression is performed between the auxiliary variable, measured from the AVHRR data set, and the target variable, measured on a sample of HRSD sites. This method has been successfully implemented by the TREES project (Mayaux et Lambin, 1995). The third method is aimed at estimating forest proportion from a sample of high resolution images : this method has been tested and implemented by FAO in the Forest Resources Assessment 1990. Moreover, if several images are acquired on each site, it is possible to estimate the deforestation rate. FAO has estimated the deforestation rate between 1980 and 1990. In the fourth approach, national mapping results are compiled at regional or global level : as for the first method, this approach is limited by the discrepancies in nomenclature, scales and dates.

### **The multi scale approach**

The second method, based on CRSD such as AVHRR data, is the only one providing some estimates of global forest resources and a location of those resources at the same time. To increase the accuracy of the estimate, many investigators suggested to use HRSD as reference data (Tucker and al., 1984; Nelson and Holben, 1988; Woodwell and al., 1987; Malingreau and Tucker, 1988; Päivinen and Witt, 1988; Malingreau and al., 1989; Stone and al., 1990; Nelson and Horning, 1993). The main advantage of this approach is the daily availability of data, and the spatial resolution (1.1 Km) which is more adapted to the scale of study. However, using coarse resolution satellite data leads to a loss of details which depends on the spatial structure of the landscape (Woodcock and Strahler, 1987; Townshend and Justice, 1988). The integration of high resolution satellite data represents a double advantage with the possibility to first validate the broad scale maps by comparing fine scale observations with AVHRR classification and second to develop a correction function with respect to proportional errors in order to estimate tropical forest areas from coarse resolution classifications.

### **Building a correction function : the lessons of the TREES project**

The Joint Research Centre TREES project is aimed at mapping and estimating tropical forest resources, detecting and monitoring changes and modeling deforestation

(Malingreau and al., 1993). A correction model for proportional errors has been tested and developed. To calibrate broad scale tropical forest area estimates, an inverse calibration model was applied (Brown, 1982; Czaplewski, 1992) by integrating in the double sampling approach a measure of the spatial fragmentation of the forest-non forest interface (Mayaux and Lambin, 1995). The inversion model can be expressed as (Eq. 1):

$$P_{F_i} = f[P_{C_i}, S_{C_i}] \quad (1)$$

where  $P_{F_i}$  is the forest proportion at fine scale,  $P_{C_i}$  the forest proportion at broad scale, and  $S_{C_i}$  a descriptor of the spatial pattern of the class  $i$  at coarse resolution. Several tests have shown that the Matheron index (Kleinn et al., 1993) is a good descriptor of the spatial structure in AVHRR classifications. It is defined as (Eq. 2) :

$$M = \frac{\text{number of runs between forest and other cover type pixels}}{\sqrt{(\text{number of forest pixels}) \cdot (\text{total number of pixels})}} \quad (2)$$

The two-step correction model for aggregation errors, in which forest cover proportion at fine and coarse resolution as well as the fragmentation index are calculated in 13x13 AVHRR pixels blocks, was developed with two major regression phases : first, a linear regression was performed within each equal-size subset of similar spatial pattern between forest cover proportions at fine and coarse resolution, and second, a regression was performed between the fragmentation measure and the intercept and the slope of the previous regressions. This calibration procedure improved by 25.5 % the reliability of the retrieval of forest cover proportions from coarse resolution data by comparison to a simple correction function relating directly proportions at coarse and fine resolutions (Mayaux and Lambin, 1995). Further improvements are in progress, with the introduction of spectral mixture models at coarse resolution scale. The integration of spatial information in the mixed pixel estimator improved significantly the reliability of the model (Mayaux and Lambin, 1996).

## **FOREST CHANGE DETECTION**

### **Digital change detection methods**

Forest ecosystems are in continuous evolution. Changes can be defined as "an alteration in the surface components of the vegetation cover" (Milne, 1988) or as "a spectral/spatial movement of a vegetation entity over time" (Lund, 1983). The intensity of change can be abrupt (clear cutting, deforestation, and forest fires) or diffuse and gradual (stands growth). Various classifications of forest change have been proposed by Aldrich (1975) who suggested nine general forest disturbance classes, by Colwell (1980) who proposed a more hierarchical framework, and by Häme (1986) who took a more mechanistic view of the problem. Ecological aspects have been then developed by Hobbs (1990), and Khorram (1994) who concentrated more on the spatial environment in which the change occurs.

Critical issues in forest change detection are the appropriate selection of images, the choice of sensors, the change categories and the change detection algorithms (Coppin, 1994). Moreover, the change interval length, e.g. the temporal resolution, is a key aspect in change detection. The accuracy of forest cover change information is closely related to the maximization of the signal-to noise ratio. When this ratio is not optimum, unreal

changes may be detected, and are caused by differences in scattering conditions, differences in water content in the atmosphere, variations in the solar azimuth angles, and sensors calibration inconsistencies (Hade, 1988). Most digital change detection methods are based on per-pixel classifiers (Teuber, 1990). But the decision rules, which comprise the complete range of pattern recognition techniques, constitute the most sensitive aspect in any change detection model. Numerous detection algorithms have been developed during the past years. They can be grouped into two main categories : change measurement methods, and classification schemes (Malila, 1980; Pilon and al., 1987). It is also possible to separate the simultaneous analysis of multitemporal data versus comparative analysis of independent single date classifications (Singh, 1989). Change detection methods may be characterized by two main phases : first the data transformation procedure, and second the analysis techniques applied to delineate the changed areas (Nelson, 1983). The transformation procedure can be no transformation (raw data), images differencing, bands ratioing, vegetation index differencing, regression between bands, multitemporal linear data transformation (principal component analysis and tasseled cap), change vector analysis, and comparison of single date classifications. The analysis techniques used to detect change are usually deriving from thresholding, supervised or unsupervised classification, spatial analysis, or layered spectral/temporal analysis. For a review of change detection techniques, see Singh (1989) and Coppin (1994).

### **The multi scale approach in change detection**

#### **How to conciliate broad scale and fine scale observations**

Most of the detection change studies have used high resolution satellite data since the changes are easier to detect at fine scale. At broad scale, change vector analysis techniques have been applied to characterize vegetation changes in Africa (Lambin and Strahler, 1994). Very few studies have been carried out using several resolution satellite data. As far as vegetation monitoring at global scale is concerned, there is a need to get observations at different scales both in temporal and spatial terms (Malingreau and Belward, 1992). A relationship between forest clearing rate based on MSS-TM analysis and fire activity based on AVHRR interpretation was studied in Mato Grosso, Brazil (Nelson and al., 1987). The results suggested to use AVHRR data as a stratification tool rather than as a auxiliary variable in the sampling design. Townshend and Justice (1995) have looked into the spatial frequencies of two MSS single date images and of the change image obtained by subtracting the NDVI. Considerable variability in the spatial frequency was found on the two images, with a spatial resolution ranging from 125 m to 64 Km. But no relationship was detected between the spatial frequencies of the two single date images and the change image.

To estimate changes, a simple approach can be applied consisting in subtracting one map (or classification) from another. But the bias and the resulting variance are unknown and might be substantial. Statistical procedures were tested for estimating change from thematic maps available for the same region at two times (Van Deusen, 1994). Subsampling was used to obtain estimates of the true class of some pixels at each of the two times. Simulation results verified that the procedures produce unbiased estimates and that associated variance estimates are reliable. Van Deusen pointed out that this statistical approach could be applied on low resolution thematic maps by using

subsamples of high resolution data to refine the true class proportions. In a global forest monitoring exercise, Beltz and al. (1992) recommended to use multi-stage sampling with sampling units proportional to forest cover or deforestation, and sampling with partial replacement.

### **A correction function for change detection**

As for the estimate of forest areas, an inverse correction model can be built to refine the estimate of forest cover changes. A simple model can be built by relating the estimate of change at broad scale with the estimate of change at fine scale (Eq. 3) :

$$\Delta P_{F_i} = f[\Delta P_{C_i}] \quad (3)$$

where  $\Delta P_{F_i}$  and  $\Delta P_{C_i}$  are the estimates of changes at fine scale and coarse scale of class  $i$  between time  $t_1$  and  $t_2$ . The introduction of the spatial structure in the model would lead to the following equation (Eq. 4) :

$$\Delta P_{F_i} = f[\Delta P_{C_i}, S_{C_i}(t_1), S_{C_i}(t_2)] \quad (4)$$

where  $S_{C_i}(t_1)$  and  $S_{C_i}(t_2)$  are the measurements of the forest spatial patterns at time  $t_1$  and  $t_2$ .  $S_{C_i}(t_1)$  characterizes the spatial structure at the initial stage, e.g. at the beginning of the time interval during which the change in forest proportion will be estimated : this a priori knowledge of the spatial arrangement of forest cover can be considered as an indicator of future deforestation. A fragmented pattern with many forest-non forest contacts may be more sensitive to deforestation than a continuous block of forest cover. However, this indicator of future deforestation and the magnitude of change are very much depending on the type of deforestation, e.g. logging activities, clear cutting, shifting cultivations...etc. On the contrary,  $S_{C_i}(t_2)$  characterizes the spatial structure at the end of the interval, and thus can be considered as an indicator of past deforestation activities. When CRSD data are not available at time  $t_1$  and  $t_2$ , the correction model can be simplified as follows (Eq. 5), with  $t_j$  being between  $t_1$  and  $t_2$  :

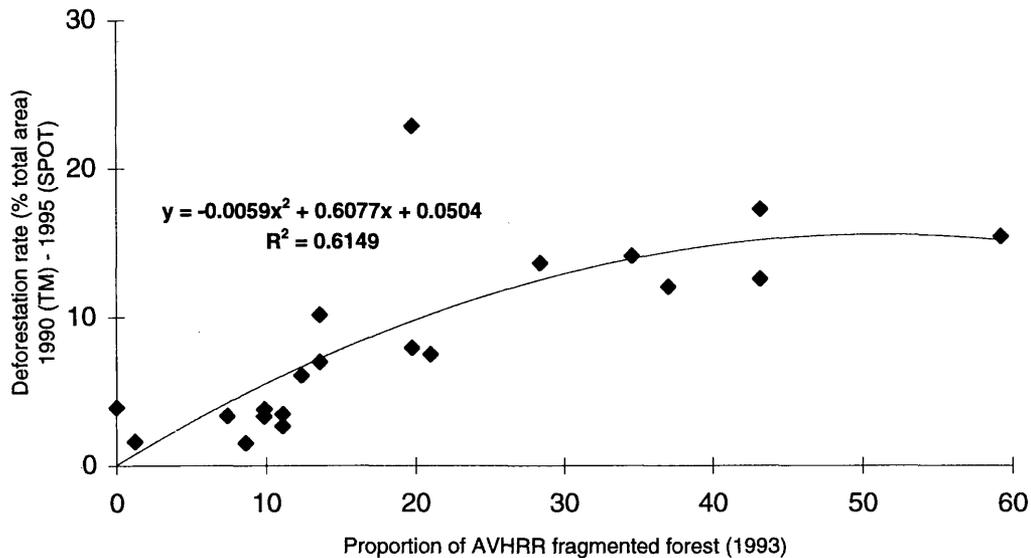
$$\Delta P_{F_i} = f[P_{C_i}(t_j), S_{C_i}(t_j)] \quad (5)$$

### **Preliminary results on a test site**

A study was carried out on a test area located in Vietnam where deforestation is very active, and field observations were available. An AVHRR classification from 1992-93 and a TM scene (124/52 from December 30, 1990) were provided by the TREES project. A SPOT image was purchased (KJ 277-326 from February 25, 1995). The acquisition periods of the TM and SPOT scenes correspond to the dry season. Geometric corrections were performed between all satellite data with a common projection system. The total RMS between the SPOT scene and the TM scene was 0.13 pixel. The RMS obtained between AVHRR data and the TM scene was 0.11 AVHRR pixel. A multitemporal unsupervised classification was performed using the three bands of SPOT and the first five bands of TM. The classes were then grouped into four classes : unchanged forest, unchanged non forest, deforested areas (between 1990 and 1995), and clouds or shadows. The AVHRR classification is giving three main classes : dense forest

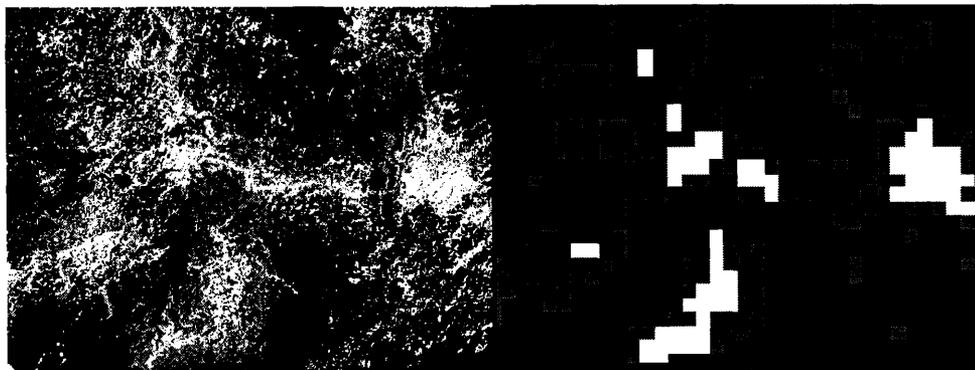
(forest cover percentage over 70 %), fragmented forest (forest cover percentage between 40 and 70 %), and non forest (forest cover percentage below 40 %). Forest cover proportions were calculated in 9x9 blocks of AVHRR pixels.

*Figure 1 . Relationship between the deforestation rate at fine scale and the proportion of AVHRR fragmented forest.*



The relationship between the deforestation rate measured at fine scale and the proportion of AVHRR fragmented forest (figure 1) shows a rather good correlation (coefficient of determination is 0.61). It can be noticed that the deforestation rate increases with the proportion of fragmented forest, but this trend stops at a proportion of 50 %. A similar trend is observed with the Matheron index, with a less good correlation with the deforestation rate ( $r^2 = 0.33$ ). On the contrary, when AVHRR dense forest increases, the deforestation rate decreases. It can be concluded that the deforestation rate is in this case governed by the fragmentation pattern observed at broad scale. The AVHRR classification and the TM/SPOT classification are presented in figure 2.

*Figure 2. Left : multitemporal classification TM(1990)-SPOT(1995) with unchanged forest (black), unchanged non forest (white) and deforested areas (grey). Right : AVHRR classification (1993) with dense forest (black), fragmented forest (grey) and non forest (white).*



It can be concluded that there is a consistent relationship between the estimate of deforestation at fine scale and forest cover parameters at broad scale. The AVHRR fragmented forest class is already containing some information of the degree of fragmentation within the pixel. The fragmentation pattern measured with the Matheron index in a block of pixels is likely to improve the correction model. However, more HRSD sites are needed to test the model and its real performance. Moreover, it is suggested to develop a more complete model integrating the estimate of forest cover change at broad scale (eq. 4). For this purpose, a more complete set of AVHRR data is required. Phase 2 of the TREES project which has just been launched should focus on global forest cover change with the acquisition and processing of new AVHRR data set. It can be also pointed out that the correction model should pave the way for a global deforestation model with the integration of auxiliary parameters such as socio-economic data, population figures, infrastructure, and local environment conditions (topography, soils...). The Tropical Forest Information System (TFIS) set up by the TREES project will be the basis for reaching this final goal.

## REFERENCES

- Achard F., and D'Souza G., 1994, Collection and processing of NOAA AVHRR 1 km resolution data for tropical forest resource assessment, TREES Series A: Technical document n°2, EUR 16055 EN, European Commission, Luxembourg, 58p.
- Beltz R., Evans D., Czaplowski R.L. and Van Deusen P., 1992, National forest area and rates of change estimates using satellite data, Report of the UNEP/FAO Expert Consultation on Environmental Parameters in Future Global Forest Assessments.
- Belward A.S., 1992, Spatial attributes of AVHRR imagery for environmental monitoring, *Int. J. Remote Sensing*, vol. 13, N°2, 193-208.
- Chomentowski W., Salas B. and Skole D., 1994, Landsat Pathfinder project advances deforestation mapping, *GIS World* 7(4): 34-38.
- Coppin P.R. and Bauer M.E., July 1994, Processing of multitemporal landsat TM imagery to optimize extraction of forest cover change features, *IEEE Transactions on geoscience and remote sensing*, Vol. 32, N°4.
- Coppin P.R., September 1994, Digital change detection in forest ecosystems: where are we and where are we going?, *ISPRS, Proceedings "Resource and environmental monitoring"*, Brazil.
- Czaplowski R.L. and Catts G.P., 1992, Calibration of remotely sensed proportion or area estimates for misclassification error, *Remote Sens. Environ.* 39:29-43.
- Czaplowski R.L., Analysis of alternative sample survey designs, FAO 1991.
- Food and Agriculture Organization, 1993, Forest resources assessment 1990: tropical countries, FAO Forestry Paper 112, Rome, 61pp.
- Jeanjean H., Malingreau J.P. and Achard F., 1994, Tropical forest fragmentation: typology and characterisation, in *Proceedings of the European Symposium on Satellite Remote Sensing*, 26-30 September, Rome.
- Klein C., Dees M. and Pelz D.R., 1993, Sampling aspects in the TREES project: global inventory of tropical forests, Final report to the Joint Research Center, Contract n° 5014-92-10 ED ISP D, Universitat Freiburg, 36pp.
- Lambin E.F. and Strahler A.H., 1994, Change-vector analysis in multitemporal space: a tool to detect and categorize land-cover change processes using high temporal-resolution satellite data, *Remote Sens. Environ.* 48:231-244.
- Lambin E.F. and Strahler A.H., 1994, Indicators of land-cover change for change-vector analysis in multitemporal space at coarse spatial scales, *Int. J. Remote Sensing*, Vol. 15, N°10, 2099-2119.
- Malingreau J.P. and Belward A.S., 1992, Scale considerations in vegetation monitoring using AVHRR data, *Int. J. Remote Sensing*, Vol. 13, N°12, 2289-2307.

- Malingreau J.P. and Tucker C.J., 1988, Large scale deforestation in the Southeastern Amazon Basin of Brazil, *AMBIO* 17:49-55.
- Malingreau J.P. and Tucker C.J., 1990, Cover, Ranching in the Amazon Basin, Large-scale changes observed by AVHRR, *Int. J. Remote Sensing*, Vol. 11, N°2, 187-189.
- Malingreau J.P., Achard F., D'Souza G. et al., 1995, AVHRR for global tropical forest monitoring: the lessons of the TREES project, *Remote Sensing Reviews* 12: 29-40.
- Malingreau J.P., Tucker C.J. and Laporte N., 1989, AVHRR for monitoring global tropical deforestation, *Int. J. Remote Sensing* 10: 855-867.
- Mayaux P. and Lambin E.F., 1995, Estimation of tropical forest area from coarse spatial resolution data: a two-step correction function for proportionnal errors due to spatial aggregation, *Remote Sensing Environ.* 53: 1-15, Elsevier Science Inc.
- Mayaux P. and Lambin E.F., September 1995, Improvements of the calibration of coarse resolution tropical forest area estimates with spatial texture measures, Manuscript submitted to *Remote Sensing of Environment*.
- Moody A. and Curtis E. Woodcock, May 1994, Scale-dependent errors in the estimation of land-cover proportions: implications for global land-cover datasets, *Photogrammetric Engineering & Remote sensing*, Vol. 60, N°5, 585-594pp.
- Nelson R. and Holben Brent, 1986, Identifying deforestation in Brazil using multiresolution satellite data, *Int. J. Remote Sensing*, vol. 7, N°3, 429-448.
- Nelson R., February 1989, Regression and ratio estimators to integrate AVHRR and MSS data, *Remote Sens. Environ.* 30:201-216.
- Nelson R., Horning N. and Stone T.A., 1987, Determining the rate of forest conversion in Mato Grosso, Brazil, using Landsat MSS and AVHRR data, *Int. J. Remote Sensing*, Vol. 8, N°12, 1767-1784.
- Päivinen R., 1994, IUFRO International guidelines for forest monitoring, Directrices internacionales de IUFRO para la monitorizacion de los recursos forestales, *IUFRO World series*, Vol. 5, ISSN 1016-3263.
- Sader S.A. and Winne J.C., 1992, RGB-NDVI colour composites for visualizing forest change dynamics, *Int. J. Remote Sensing*, Vol. 13, N°16, 3055-3067.
- Singh A., 1989, Digital change detection techniques using remotely-sensed data, *Int. J. Remote Sensing*, Vol. 10, N°6, 989-1003.
- Stone T.A. and Schlesinger P., June 1990, Monitoring deforestation in the tropics with NOAA AVHRR and Landsat data.
- Townshend J.R.G. and Justice C.O., 1988, Selecting the spatial resolution of satellite sensors required for global monitoring of land transformations, *Int. J. Remote Sensing*, Vol. 9, N°2, 187-236.
- Townshend J.R.G. and Justice C.O., 1995, Spatial variability of images and the monitoring of changes in the Normalized Difference Vegetation Index, *Int. J. Remote Sensing*, Vol. 16, N°12, 2187-2195.
- Van Deusen P.C., 1994, Correcting bias in change thematic maps, *Remote Sens. Environ.* 50:67-73.
- Woodcock C.E. and Strahler A.H., 1987, The factor of scale in remote sensing, *Remote Sens. Environ.* 21:311-332.
- Zhu Z., 1994, Forest density mapping in the lower 48 States: a regression procedure, U.S. Department of Agriculture, Forest Service, Southern Forest Experiment Station, New Orleans, LA, Research Paper SO-280, 9pp.

# Optimum Area Sampling Frame Using High Resolution Satellite Images with Operational Objective : How to Conciliate Statistical Requirements and Practical Aspects?

Hélène de Boissezon<sup>1</sup>, Hervé Jeanjean<sup>2</sup>

---

**Abstract:** If high resolution satellite images can be used for measuring land use characteristics, it is most often impossible to envisage an exhaustive coverage of large areas. Sample survey methods represent a feasible alternative when assessing land use surface. In many cases, classical sampling survey techniques can be used. However, when spatial correlation is observed between sites, geostatistics principles could be introduced into the inventory design. Using classical statistical techniques, the target variable has to be estimated from the elementary measurements on the sample sites through an extrapolation model. When designing such statistical process, many issues have to be addressed, such as the characteristics of the population to be considered, the sampling design itself, the extrapolation model to be applied on the whole study area, and the sampling error. Practical considerations and constraints should not be underestimated in any sample survey. In this paper, statistical issues on sampling frame design based on high resolution satellite images are addressed with emphasis on practical implementation. Sampling survey designs of some operational remote sensing projects in the world are reviewed, such as FAO/FRA 90, and the European MARS, TREES, FIRS programmes. Some recommendations are then proposed for setting up sound sampling schemes.

---

## INTRODUCTION

In Earth observation, the analysis and monitoring of land surface characteristics such as proportion of crops or forest types are often requiring high resolution satellite images. Although some programmes are using this type of data on a wall to wall basis, it becomes difficult to assess large areas with a full

---

<sup>1</sup> Remote sensing expert in agronomic statistics, SCOT Conseil, 1 rue Hermès, F-31526 Ramonville France

<sup>2</sup> Remote sensing expert in forestry, SCOT Conseil, 1 rue Hermès, F-31526 Ramonville France

coverage, because of the high cost induced and the huge amount of data to be processed. A sampling scheme, based on sound statistical concepts, represents a feasible method. Some operational regional and international programmes have already developed and implemented such procedures. However, the sampling scheme has been sometimes designed *a posteriori*, e.g. the samples have not been selected and located under rigorous rules at the beginning. Moreover, practical constraints can be either overestimated or underestimated, whether they are given too much or too little consideration. This situation results in problems in final accuracy and in implementing an extrapolation model. This paper is addressing the optimization of an optimum area sampling frame design : consideration will be given to both statistical requirements and practical constraints.

## **RECENT ADVANCES IN SAMPLING DESIGNS**

Few investigations have been carried out on the most adapted statistical procedures for covering large areas with a sample of high resolution images. However, the literature is rather rich in research studies concerning small areas within satellite images, where sampling schemes can be tested and evaluated (the true variance can be calculated, and the sampling can be repeated as many times as wishes).

### **The geostatistical approach**

Geostatistical concepts constitute an alternative to classical sampling techniques. They assume that the sample units are not independent, and that the autocorrelation can be used in the model. Geostatistics requires a systematic sampling rather than a random sampling. The semivariogram is the basic tool in geostatistics. In remote sensing images, strong correlation has been observed between landscapes (Ramstein and Raffy, 1989; Webster and al., 1989; Gohin and Langlois, 1993; Wood and al., 1994). McGwire and al. (1993) has shown that the spatial correlation is increasing with the sample unit size. Atkinson (1995) confirms this result and has pointed out that the gain of a systematic sampling by comparison to a random sampling is increasing with the size of the sample unit. A minimum distance between sample sites can reduce the effects of the spatial correlation (the sampling units become less dependent), and thus can improve the efficiency of the scheme. According to McGwire and al. (1993), geostatistics may be tested and applied in wide vegetation mapping and monitoring projects. The studies carried out by Jupp and al. (1988) constitute a good starting point for developing such methods. In spite of this suggestions, no complete study has proved the applicability of geostatistical concepts when sampling large areas (country to continent scale) with high resolution images. In particular, the spatial correlation between SPOT or TM scenes selected on a systematic basis has still to be proved. The main bottleneck for evaluating the usefulness of geostatistical bases

is related to the important means (number of satellite scenes and processing time) which would be required for building and testing a semivariogram. Therefore, classical statistical concepts are generally still prevailing since their application rules and their limits are well known.

### **Gain of systematic sampling**

Previous studies have shown that systematic sampling is becoming more efficient when the sample size or the sampling intensity increases (Atkinson, 1995). In fact, as the sample size increases, the sampling units are likely to be more and more correlated since they are closer to each other. However, it is always recommended to carefully examine the spatial behaviour of the variable to estimate. If the variable is following a periodicity (sequence of ridges and valleys for instance), a bias is likely to be introduced. If several simulations confirmed the gain of systematic sampling (Cochran, 1981; Brion and Fournier, 1995), the magnitude of the gain is varying and is depending on the spatial correlation (Dunn and Harrison, 1993).

### **Multi phase/stage sampling**

When the population to be assessed is not well known, e.g. when stratification is not possible, it may be convenient to proceed to a multiple phase sampling. In remote sensing, a double phase sampling with regression has been tested or recommended in many studies, whether between high resolution images and ground measurements (Smiatek, 1995) or between coarse resolution satellite images and high resolution satellite images (Mayaux and Lambin, 1995; Kleinn and al., 1993; Päivinen and Pitkänen, 1992; Iverson and al., 1989; Nelson and Holben, 1986). Multistage sampling is a feasible alternative. In that case, the total variance is often mainly due to the first stage.

### **Gain of stratification**

The *a priori* stratification consists in partitioning the population with criteria well correlated to the variable to estimate. The final gain is depending on this relationship. The *a posteriori* stratification, which consists in grouping sampling units in ensembles defined after the sampling scheme, may provide good results (Cochran, 1981; Dunn and Harrison, 1983).

## **EXPERIENCE OF SOME LARGE SCALE REMOTE SENSING PROJECT**

Several operational remote sensing projects have set up sound sampling schemes based on high resolution satellite images as sampling units. The activity B of the MARS project (Monitoring Agriculture by Remote Sensing) is an European

programme carried out by the Joint Research Centre at the Institute of Remote Sensing Applications and is aimed at providing fast estimates of crops areas and yields to DGVI in Brussels. A total of 60 sampling sites (40x40 Km<sup>2</sup>) have been located after stratification, which corresponds to 6 % sampling intensity. The allocation of sites is proportional to the agricultural lands in the country, and is coinciding with SPOT scenes (KJ). A weighting procedure is then employed to derive the extrapolation model. The MERA 95 project (MARS extension and Environment Related Applications) is an extension of MARS to Central and Eastern Europe. The sampling design has been slightly modified, with no stratification, and 40 sites have been selected among a systematic sample base, with proportionality to the agricultural activity. The FAO Forest Resources Assessment 1990, following recommendations given by statisticians (Czaplewski, 1991) is using a sample of 117 TM sites located randomly in the tropical belt (FAO, 1993). The variables to be estimated are the forest areas and the change of forest cover in the last 10 years. The sampling intensity is 10 %, and a stratification has been performed using ecological criteria. The sample size has been fixed with an objective of 5% of the standard error. The deforestation rate was estimated with an error of 12.5 %, which can be explained by the high coefficient of variation of deforestation. The FIRS project (Forest Information by Remote Sensing) is another activity of JRC, Italy. One of its fundamental activities was to stratify Europe into 115 homogeneous strata with three main criteria : proportion of forest cover, species composition, and timber volume. The strata are grouped into 6 statistical regions, where 223 sites have been selected with unaligned systematic sampling. The sampling intensity is 3 % of the total area, or 10 % of the forest lands. The TREES project (Tropical Ecosystem Environment observation by Satellites) is using a double sampling scheme with regression between AVHRR derived forest cover proportion as the auxiliary variable and TM derived forest cover proportion as the target variable. The sampling unit is corresponding to a block of 11x11 AVHRR pixels. A total of 1121 units have been selected, each TM scene generating more than 200 units (cluster sampling). A regression is performed between the forest proportions estimated at TM and AVHRR levels.

## **DESIGNING AN OPTIMUM SCHEME**

### **Constraints**

The constraints in designing an operational sampling scheme are related to the satellite itself (availability of data, acquisition delays, repetitivity for multitemporal analysis...), the operability of the system (fulfilment of the requirements and technical feasibility), efficiency and cost-effectiveness (the measurement error induced by remote sensing techniques must be compared and added to the sampling error), and the client requirements (terms of reference and required precision at which level of confidence).

### Definition of a sound sampling scheme using HRSD

At a very first stage, it is necessary to identify which land cover parameters are to be analysed. The sampling scheme will depend on the characteristics of the land cover type under investigation, e. g. its surface proportion in the entire domain (study area), the size of the elementary components of this land cover type (size of agricultural plots for instance), its spatial distribution (presence or absence in the different strata of the study area), its occurring frequency (proportion of the land cover type in the strata), and its temporal variation (crops cycle for instance). Based on high resolution satellite images, the variable to estimate with satisfactory precision can be the proportion of the land cover type, e.g. the proportion of agricultural areas or the forest areas. It can be also the variation of proportion with a multitemporal data set. The *a priori* knowledge of the domain is a critical issue for setting up a sound sampling scheme. The sample must be as representative as possible of the entire population. If limited knowledge is available, multiphase sampling is recommended. In table 1, sampling schemes are proposed with respect to the characteristics of the item.

**Table 1.-Designing a sampling scheme based on the characteristics of the item  
(Cochran, 1981).**

charateristics of item	type of sampling needed
1. widespread throughout the region, occurring with reasonable frequency in all parts.	a general survey with low sampling intensity.
2. widespread throughout the region but with low frequency.	a general survey, but with a higher sampling ratio.
3. occurring with reasonable frequency in most parts of the region, but with more sporadic distribution, being absent in some parts and highly concentrated in others.	for best results, a stratified sample with different intensities in different parts of the region. Can sometimes be included in a general survey with supplementary sampling.
4. distribution very sporadic or concentrated in a small part of the region.	not suitable for a general survey. Requires a sample geared to its distribution.

The sampling type has to be examined when using satellite images. Indeed, the measurements are not made on a single point, but within an area (SPOT scene for instance). The results may be affected to the scene centre (sample of points taken from an infinity of points) or to the entire area (area frame sampling with a sample taken from N sites covering the entire region). The definition of the sample base is an important step of the sampling scheme : it is, in principle, composed of all units (points or frames) constituting the entire region. When using SPOT scenes centres as the sample base on very large region, a geometric problem is raised by the narrowing of tracks towards the pole, which induces of higher density of units. The sampling scheme should take into consideration this distortion by selecting the

sample-units on a proportional basis (proportional to the distance between two orbits for instance). A stratification is then often required, mostly when the a priori proportion of the item is known with sufficient accuracy. The ideal variate for stratification is the value of the variable to estimate. In practice, other source of information is used, such as ancillary data (existing maps or statistics). The stratification process can lead to different types of strata : the strata in which the item is almost absent and in which no sample will be taken, and the strata with different proportion of the item (from low to high proportion). It is sometimes recommended to delineate special strata which correspond to other criteria (mountainous areas, swampy zones...etc). It is important to measure the total area of each stratum, and to verify that the sum is giving the total area of the inventoried region. The number of strata will be low if the *a priori* knowledge of the population is limited. Systematic sampling in two dimensions represent a good area sampling frame scheme. An alternative is the unaligned sampling (Cochran, 1981). A combination of systematic sampling (selecting the cluster) and random sampling (selecting the unit in the cluster) may also give good results. The number of sample units within each stratum may be calculated with the following equation (Eq. 1) :

$$n_j = \frac{n(S_j \sqrt{\text{Var}_j})}{\sum_j S_j \sqrt{\text{Var}_j}} \quad (1)$$

with  $n$  as the total sample size,  $n_j$  the sample size in stratum  $j$ ,  $S_j$  the size of stratum  $j$ , and  $\text{Var}_j$  the variance of stratum  $j$ . The total sample size  $n$  is determined by the required overall precision and the available budget. For optimizing the design (cost effectiveness), the Lagrange equation can be used to solve the system. In stratified sampling, the overall estimate of the mean value is weighted by the strata sizes (Eq. 2) :

$$\bar{Y} = \frac{\sum_{j=1}^J N_j \bar{Y}_j}{N} \quad (2)$$

with  $N_j$  as the stratum size,  $\bar{Y}_j$  the estimated mean value in stratum  $j$ , and  $N$  the total number of sample units. The total error of the design is a sum of the sampling error and the measurements error. When the measurement error is important, it is not worth designing a complex sampling scheme with a very high precision since the overall precision will be low.

## CONCLUSION

In sampling statistics, the estimates are always given with a confidence interval and a probability level. This means that we are not absolutely sure of the error we

have made with the sample. It is sometimes difficult to decide how much error should be tolerated by the client or the user. Moreover, starting from the user requirements and means (budget allocated for the survey), and looking into the characteristics of the parameter to be estimated, it should be pointed out that there is no single optimum sampling design. In our problem consisting in optimizing the estimate of the sample size and the location of the sample units, it can be concluded that geostatistical concepts should be further investigated before being applied in an operational project. The combination of several standard statistical principles often lead to satisfactory results (systematic sampling at first stage and random sampling at second stage for instance). The use of standard principles always simplifies the design where the error can be easily calculated. In operational projects, the use of a satellite reference grid (SPOT or TM) seems to be the best way for building up the sampling base (FAO/FRA 90 and MARS projects). However, in high latitudes, the distortion of the grid has to be examined in order to avoid an oversampling. This issue should be addressed by global survey over large areas, such as the agricultural statistics project in Russia which has just been launched under European Union support (TACIS) and in which SCOT Conseil is providing technical assistance.

#### REFERENCES

- Atkinson, P., 1994, Testing the efficiency of sampling strategies with simulated remotely sensed data, in SFPT, n°137, p12.
- Brion, P., et Fournier, P., 1995, Etude Géostat-Maroc: Partie statistique. Rapport CNES travaux INSEE, SCEES.
- Chevrou R.B., DERF Montpellier, Mars 1988, Inventaire Forestier National, Méthodes et Procédures.
- Cochran W. G., Mars 1981, Sampling techniques, third edition.
- Czaplewski R.L., Analysis of alternative sample survey designs, FAO 1991.
- Dagnelie P., 1973, Théorie et méthodes statistiques, Volume 1, Centre de Documentation de Toulouse.
- Dagnelie P., 1975, Théorie et méthodes statistiques, Volume 2, Centre de Documentation de Toulouse.
- Dunn R., Harrison A.R., 1993, Two-dimensional systematic sampling of land use, Appl. Statistics, 42, N° 4, pp. 585-601.
- Fitzpatrick-Lins K., Mars 1981, Comparison of sampling procedures and data analysis for a land-use and land-cover map, PERS vol. 47, N 3, pp. 343-351.
- Food and Agriculture Organization, 1993, Forest resources assessment 1990: tropical countries, FAO Forestry Paper 112, Rome, 61pp.
- Gallego FJ , P Vossen, JF Dallemand, V Perdigao, Sampling plans in MERA Project, MARS project, 1995.
- Houllier F., 6 Juin 1986, Echantillonnage et modélisation de la dynamique des peuplements forestiers/Application au cas de l'Inventaire Forestier National, Thèse.

- Iverson L. R., Cook E.A., Graham R. L., 1989, A technique for extrapolating and validating forest cover across large regions - Calibrating AVHRR data with TM data, *International Journal of Remote Sensing*, Vol. 10, n°11, p. 1805-1812.
- Kleinn C., Dees M., Pelz D.R., 1993, Sampling aspects in the TREES project - global inventory of tropical forests, final report, Freiburg Universität Germany, contract for JRC Ispra, Italie.
- McGwire K., Friedl M., Estes J.E., 1993, Spatial structure, sampling design and scale in remotely-sensed imagery of a California savanna woodland, *International Journal of Remote Sensing*, Vol. 14, n°11, p. 2137-2164.
- Nelson R. and Holben B., 1986, Identifying deforestation in Brazil using multiresolution satellite data, *International Journal of Remote Sensing*, vol. 7, pp. 429 - 448.
- Païvinen R. and Pitkänen J., 1992, Calibrating AVHRR data with TM data for tropical forest cover assessment, IUFRO S 4.02.05 Wacharakitti international workshop "Remote sensing and permanent plot techniques for world forest monitoring", Pattaya Thailand.
- Raffy M., Lacaze B., Rambal S. and Winkel T., August 1994, Identifying spatial patterns of Mediterranean landscapes from geostatistical analysis of remotely-sensed data, *International Journal of Remote Sensing*, Volume 15, Number 12, Special issue: scaling in remote sensing.
- Raffy M., Puech C., August 1994, Thresholds of homogeneity in targets in the landscape. Relationship with remote sensing, *International Journal of Remote Sensing*, Volume 15, Number 12, Special issue: scaling in remote sensing.
- SCOT CONSEIL & GAF for JRC EMAP Unit, 1995, FIRS Project, Regionalization and stratification of European forest ecosystems, final report.
- SCOT CONSEIL, CCE, 1994, Documentation Action IV, Estimations rapides des superficies et des productions potentielles au niveau européen, Volume I - Document de synthèse.
- SCOT CONSEIL, CCE, 1994, Documentation Action IV, Estimations rapides des superficies et des productions potentielles au niveau européen, Volume III - Méthodologie.
- Smiatek G., 1995, Sampling thematic mapper imagery for land use data, *Remote Sensing of Environment*, 52:116-121.

# The Use of AVHRR Satellite Imagery to Monitor Boreal Ecosystem Forest Fires

Donald R. Cahoon, Jr.<sup>1</sup> and Brian J. Stocks<sup>2</sup>

---

**Abstract** - Forest fires are an integral part of the natural forces that shape the composition and evolution of the boreal forest. These fires, often very intense, commonly grow to sizes large enough for satellite monitoring using the National Oceanic and Atmospheric Administration (NOAA) Advanced Very High Resolution Radiometer (AVHRR) instrument. The AVHRR instrument acquires 1-km and 4-km resolution (at nadir) satellite imagery in five spectral bands. Multispectral methods have been developed, using the AVHRR imagery, to estimate the size and growth rate of burned forest. The AVHRR burned area estimates have been evaluated and the error determined for both the 1-km and 4-km imagery. Applying these methods, the spatial distribution and extent of forest fires has been mapped for the 1987 and 1992 Russian boreal forest fire season. Due to a strong contrast in the total area burned in each of the 2 years, the large interannual variability of burned forest is demonstrated.

---

## INTRODUCTION

The boreal ecosystem is a forested belt, mostly comprised of conifer trees, that stretches across the Northern Hemisphere's circumpolar countries. The boreal forest covers over 12 million km<sup>2</sup> of the Earth's surface which is about 10% of the total land surface area. In terms of areal extent, the boreal forest ranks about fifth compared to all terrestrial biomes, but ranks second only to tropical forests in total plant mass (Whittaker and Likens 1975). Given the ecosystem's size and total available plant mass it is not surprising that the World's boreal forests contain almost 40% of all terrestrial carbon (Kasischke et al. 1995a). This means that the boreal forest must play an extremely important role in the global carbon budget and it is important to characterize any disturbance and process that perturbs the carbon cycle within this ecosystem. Fire is one such disturbance.

Because of the enormous geographical scale for which the boreal forest extends, a trade-off between satellite resolution, accuracy of estimating the acreage burned, and frequency of satellite coverage has been made. In reality, there are three satellite systems that could potentially be used to monitor forest fires. Landsat and SPOT quickly come to mind because of their high spatial resolution, but do not offer the repeat coverage needed to compensate for cloud cover

---

<sup>1</sup> Research Scientist, NASA/Langley Research Center, Hampton, VA.

<sup>2</sup> Senior Research Scientist, Canadian Forest Service, Sault. Ste. Marie, ONT, Canada.

obscuration of the land surface. Also, the amount of data and cost of enough imagery to cover the entire boreal region would be unrealistic. Rather, we have chosen to use AVHRR imagery because it offers a reasonable resolution at a low cost, and it provides multiple overpasses each day. After an evaluation of the AVHRR imagery, we have found that estimates of the surface area can be within a few percent. This error is considered reasonable for the large geographical scale we are evaluating as long as consistent results can be obtained.

To develop consistent surface area estimation results, we have developed a methodology for processing the AVHRR imagery. The following sections will describe the AVHRR instrument and archived imagery, the burned area estimation methodology and error analysis, and describe some of our ongoing research on boreal forest fires.

### AVHRR INSTRUMENT

The AVHRR instrument has flown on several National Oceanic and Atmospheric Administration (NOAA) sun-synchronous polar-orbiting meteorological satellites since 1978 and provides a nearly continuous archived data set. These satellites have flown in a morning and afternoon configuration which provides four overpasses each day in the low to mid-latitudes, and upwards toward 8 overpasses in the higher latitudes of the boreal forest. The AVHRR instrument is a scanning imager which scans in the across track direction and acquires data in five spectral bands (table 1). The swath width is just over 2800 km given a nominal orbital altitude of 830 km and a total scan angle of 110.8°. Each cross track scan provides 2048 observations (pixels) with an instantaneous field-of-view (IFOV) of approximately 1.4 mrad. The precise IFOV for each channel varies slightly and is included in table 1.

**Table 1. - AVHRR channel characteristics.**

Channel	Wavelength (microns)	Cross track IFOV (mrad)	Along track IFOV (mrad)
1	0.58 - 0.68	1.443	1.433
2	0.725 - 1.10	1.433	1.423
3	3.55 - 3.93	1.423	1.423
4	10.3 - 11.3	1.280	1.463
5	11.5 - 12.5	1.320	1.310

Channel selection of the AVHRR instrument was driven by operational meteorological and oceanographic requirements. However, over time the fire community has found ways to apply the channel selection to its own research. In summary, channel 1 and channel 2 can be used to differentiate smoke from clouds

(Chung and Le 1984). Channel 2, because of its sensitivity to the reflectance from vegetated surfaces, provides tremendous contrast between burn and unburned areas (Cahoon et al. 1992; Cahoon et al. 1994; Kasischke et al. 1995a). Channel 3 has long been demonstrated to be sensitive to hot temperature sources (Matson and Dozier 1981; Scorer 1987) and is useful for mapping active forest fires (Cahoon et al. 1994; Stocks et al. 1996). Channels 4 and 5 are both similar in that they are low in contrast when calibrated, but provide a basis for distinguishing between burned areas and water in mid-afternoon imagery (Cahoon et al. 1994).

The imagery is archived in either of two formats, local area coverage (LAC) or global area coverage (GAC). The LAC format is the full 1-km (at nadir) resolution product and is recorded onboard the spacecraft and downlinked to specific sites. The AVHRR 1-km data are also continuously broadcast to any ground-station within the satellite's horizon; which is referred to as high resolution picture transmission format (HRPT) data. In the archives HRPT and LAC data are the same. Due to a limited amount of onboard memory for recording LAC data, a 4/15 subsampled product is continuously recorded for broadcast to selected ground stations. This format is the global area coverage (GAC) and is frequently referred to as the 4 km imagery. The GAC subsampling scheme is shown in figure 1.

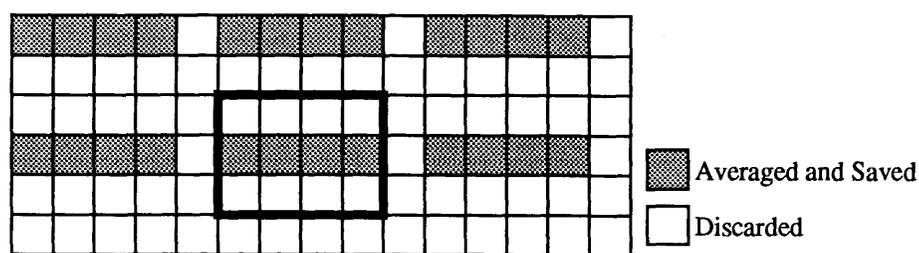


Figure 1. - GAC subsampling scheme. Each row is a portion of a scan line. Each box is one image pixel. The dark box represents a GAC "4 km" pixel.

## BURNED AREA ESTIMATION METHODOLOGY

The surface area of burn scars is estimated using either LAC or GAC archived imagery. In a previous study, Cahoon et al. (1992) demonstrated that, given the large fire sizes in the boreal system that GAC imagery is sufficient for estimating the area. In the following section we will examine the surface area estimation error using both the LAC and GAC imagery. This section will lay out the methodology through which we process the AVHRR imagery.

Because of the intermittent availability of clear-sky imagery, which negates the possibility of observing the entire study region at one time, a clear-sky mosaic is created. From this mosaic, all burn scars can be observed in a single image, making the area-burned estimate simpler and alleviating the concerns of double counting or missing any burned area. The development of the clear-sky mosaic is

a several-step process which will be described with the use of matrix notation. An individual multispectral (3-banded) AVHRR scene is represented by matrix  $D^n$  and the mosaic of all the individual AVHRR scenes is matrix  $M$ , where

$$D^n = \begin{bmatrix} d_{11} & d_{12} & \dots & d_{1j} \\ d_{21} & d_{22} & \dots & d_{2j} \\ \dots & \dots & \dots & \dots \\ d_{i1} & d_{i2} & \dots & d_{ij} \end{bmatrix}; M = \begin{bmatrix} m_{11} & m_{12} & \dots & m_{1j} \\ m_{21} & m_{22} & \dots & m_{2j} \\ \dots & \dots & \dots & \dots \\ m_{i1} & m_{i2} & \dots & m_{ij} \end{bmatrix} \quad (1)$$

$n$  is the total number of AVHRR scenes;  $i$  and  $j$  define the number of elements,  $d$  and  $m$ , in the matrices. Each pixel in the AVHRR image is corrected for brightness changes due to the solar zenith angle. Using the NOAA-supplied navigation data, each element  $d_{ij}$  is the result of resampling the original NOAA scene into a cylindrical-equidistant map projection. Since each AVHRR scene is resampled, for every  $D^n$  the elements  $d_{ij}$  represent the same geographical region and are a function of latitude ( $\phi$ ) and longitude ( $\lambda$ ). The mosaic image ( $M$ ) is the same size and spatially represents the same geographical area as the AVHRR scenes  $D^n$ . Then each element of both  $M$  and  $D^n$  can be compared, with the results of the comparison composited into a final scene ( $M'$ ). The comparison,

$$M'_{ij} = \text{MIN}(D^n_{ij}, M_{ij}) \quad (2)$$

tests for the minimum between the mosaicked image ( $M$ ) and the AVHRR scene ( $D^n$ ) in the AVHRR channel 2 band, and saves the elements for all channels of the minimum-valued image in the new mosaic  $M'$ . The testing is iterative, where  $M'$  is substituted for  $M$ , and continues  $n$  times through all of the scenes. Since clouds are highly reflective and the burned area is of low reflectance in the AVHRR channel 2, the burned areas are retained while the clouds are removed.

In estimating the total regional burned area, it was first necessary to classify which image pixels contained burned area. An unsupervised minimum distance classification on AVHRR channels 1, 2, and 4 was then used iteratively to isolate the scar pixels and statistically determine the edge pixels that are counted as having a burn scar. The iterative process continues until burn scars, that have been correlated with known fire activity determined from the AVHRR imagery, are isolated and falsely classified pixels have been minimized.

The resulting scar map is subjected to another processing step to determine the area of each pixel. Since a cylindrical-equidistant projection is used and each side of a pixel is parallel to both meridians and parallels, the distance along each pixel in terms of degrees of latitude and longitude is easy to determine as well as the surface area. The area of every pixel classified as a scar is then integrated to derive the total area burned.

## BURNED AREA ERROR ANALYSIS

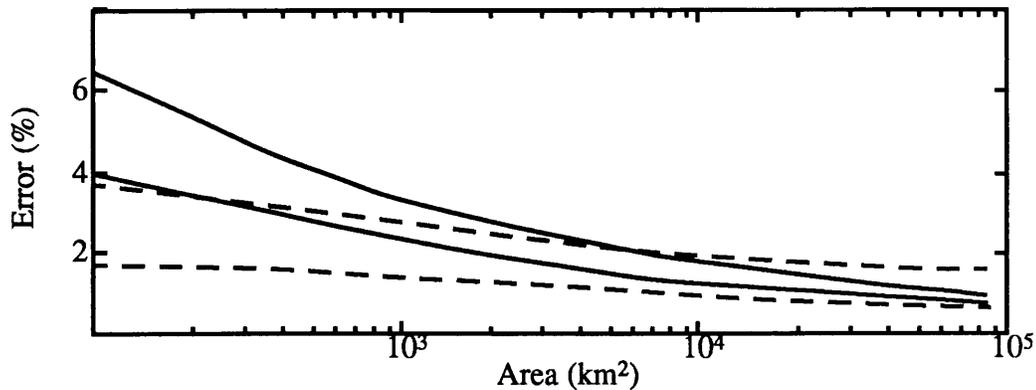
An extensive analysis has been conducted to evaluate the error of estimating the burned surface area using both the LAC and GAC imagery. The details are outlined in Cahoon et al. (1992). In brief, surface targets of known area are delineated in both the LAC and GAC imagery. These targets are a series of lakes in Canada that often contain several islands, sometimes of considerable size, and thus are not homogeneous. The heterogeneous nature of the lake targets is similar to that of burn scars which contain unburned areas within a its perimeter. The results from this analysis are in figure 2.

The mean surface area estimation error associated with using LAC imagery is lower than that of GAC imagery across the entire size range, even though for larger targets the difference is practically negligible. This comes as no surprise since we expect the analysis of higher resolution imagery to yield better results when working with small areas. However, it is interesting to note that for larger targets, the error variance for the GAC-derived estimates is less than that of the LAC analysis. This likely has a lot to do with the overlap of the AVHRR pixels (Cahoon et al. 1992). The overlap is considerable for the 1-km data and the edge of a target is defined by a gradual gradient, leaving room for interpretation as to where to define the target's edge. Much of the surface area estimate errors are due to blurred edges of targets. At least for the GAC imagery, some of this overlap has been eliminated by subsampling, leaving a better defined edge and fewer perimeter pixels to skew the analysis one way or another.

Overall, the results show that for LAC imagery a mean error of about 2% can be expected, and that 90% of all targets will fall within 4% error for targets down to 100 km<sup>2</sup>. For the GAC imagery, the mean error is about 2% worse than that of the LAC imagery for targets that are 100 km<sup>2</sup>, and drops off to being negligible when evaluating targets that are 10,000 km<sup>2</sup> or more in size. Burned area estimate errors, from both the 1988 Yellowstone Fires (Cahoon et al. 1992) and the 1987 Great Dragon Fire in China (Cahoon et al. 1994), are consistent with these results.

### RESEARCH APPLICATION: RUSSIAN BOREAL FOREST FIRES

Forest fires have been a major disturbance regime affecting the world's boreal forests for millennia. Increased development within the world's boreal zone has necessitated fire management programs designed to protect human interest and forest investment. While these programs have been largely effective in reducing unwanted fires in selected regions, forest fires continue to exert enormous influence in the world's boreal forests. With the rising concern that potential global warming will lead to an increase of fire activity within the boreal forests (Flannigan and Van Wagner 1991, Wotton and Flannigan 1993, Stocks 1993, Fosberg et al. 1996), much attention as recently been focused on the important role the boreal forest plays in the Earth's carbon budget (Kasischke et al. 1995b), and how an increase in fire activity will impact that budget and potentially provide



**Figure 2. - AVHRR-derived surface area estimation error for various target sizes. LAC errors shown by dashed line; GAC errors shown by solid line. The lower line, for both LAC and GAC, represents mean error and the upper line shows the upper error bound for 90% of all targets.**

positive feedback to global warming (Kurz et al. 1994). Given these concerns and the need for a quantitative assessment and a historical perspective of boreal fire activity, we have begun to use the satellite techniques described in this paper to assess the total area burned in the boreal forest during the last 15 years for which AVHRR data are archived.

Leading into this multiyear study, as test cases, we have completed the burned area analysis for the Russian boreal forest for the years 1987 (Cahoon et al. 1994) and 1992 (Cahoon et al. 1996). The 2 years were totally dissimilar in terms of area burned, with 1987 being a severe fire year for both Russia and the northern Chinese provinces. It was during 1987 that one of the largest recorded fires in China's history occurred (about 1.3 million ha) and large fires throughout eastern Russia were very prevalent (Cahoon et al. 1994). The 1987 analysis was conducted using over 200 GAC images. As demonstrated in the previous section, the burned surface area estimation using GAC imagery is not much greater than 6% for fire sizes down to 100 km<sup>2</sup> in size. Fortunately, in the boreal forest, fire sizes are very large with 95% of the area burned by 5% of the fires. The estimate of the total area burned during the 1987 fire season is 14.446 million ha.

In contrast, even though there was widespread fire activity in 1992, the total area burned was 1.5 million hectares. This is a relatively low annual total of burned acreage which was concurrently experienced in Alaska and Canada. The 1992 analysis used over 100 LAC images that cover the region from 30° to 180° east longitude. The boreal region was split into 30° longitude blocks. Each block was processed one at a time (Cahoon et al. 1996).

There was an obvious similarity between 1987 and 1992, widespread fire activity is primary east of the Yenisey River (about 90°N latitude). The reason for the higher amount of burned area to the east is likely due to the increase in population toward the west and the low marshy areas west of the Yenisey River. With the population increase to the west comes additional fire management pressures and better access to fires for fire suppression purposes.

## CONCLUSIONS

With limited resources and the constant need to assess spatial changes within our natural resource regimes, satellite and aircraft remote sensing is being drawn upon to provide spatially dependant data over various scales. The need to monitor and assess the post-burned size of forest fires is one such role that satellite-based observation can play. Even though it is more difficult to used space-based resources to contribute directly to fire fighting, it is not out of the question to utilize this resource to monitor fire activity and map post-burned fires over large geographical scales and remote regions.

The remoteness of the boreal ecosystem, a region where considerable travel and fire reconnaissance is accomplished by air, almost dictates the need in a cost-cutting era to assess fire activity using satellite imagery. The remoteness, combined with the current interests in developing a consistent fire record throughout the entire boreal forest, has lead us to develop the AVHRR techniques used in our two trial studies, which will be implemented as we embark on a 15-year analysis of the interannual variations of burned forest. Further evaluations will be made regarding the successfulness of detecting and estimating the size of smaller fires. Further, a test case has also been developed which demonstrates the potential use of the AVHRR imagery for monitoring fire growth rates (Stocks et al 1996).

To conduct this research, customized computer code has been developed over a number of years to implement our approach. Each segment of our software system has been carefully tested. We have begun to evaluate the potential of commercial packages for registering and mapping AVHRR images. In one prominent GIS/image processing package, we have had very limited success with registration and would not trust the resulting surface area estimates. This does lead to our concern that any commercial tools that can map fires and estimate the size of burned areas using AVHRR satellite imagery should carefully be scrutinized for reliability.

## REFERENCES

- Cahoon, Jr., D. R., B. J. Stocks, J. S. Levine, W. R. Cofer III, and C. C. Chung, 1992, Evaluation of a technique for satellite-derived estimation of biomass burning, *J. Geophys. Res.*, 97(D4): 3805-3814.
- Cahoon, Jr., D. R., B. J. Stocks, J. S. Levine, W. R. Cofer III, and J. M. Pierson, 1994, Satellite analysis of the severe 1987 forest fires in northern China and southeastern Siberia, *J. Geophys. Res.*, 99(D9): 18,627-18,638.
- Cahoon, Jr., D. R., B. J. Stocks, J. S. Levine, W. R. Cofer, and J. A. Barber, 1996, Monitoring the 1992 Forest Fires in the Boreal Ecosystem using NOAA AVHRR Satellite Imagery: Biomass Burning and Global Change, MIT Press, Cambridge, MA. (in press)
- Chung, Y. S. and H. V. Le, 1984, Detection of forest-fire smoke plumes by satellite imagery, *Atmos. Environ.*, 18(10): 2143-2151.

- Flannigan, M. D., and C. E. Van Wagner, 1991, Climate change and wildfire in Canada, Can. J. For. Res., 21: 66-72.
- Fosberg, M. A., B. J. Stocks, and T. J. Lynham, 1996, Risk analysis in strategic planning: fire and climate change in the boreal forest: Fire in Ecosystems of Boreal Eurasia, Blackwell Press, Oxford, UK. (in press)
- Kasischke, E. S., N. H. F. French, L. L. Bourgeau-Chavez, and N. L. Christensen, Jr., 1995, Estimating release of carbon from 1990 and 1991 forest fires in Alaska, J. Geophys. Res., 100: 2941-2951.
- Kasischke, E. S., N. L. Christensen, and B. J. Stocks, 1995, Fire, global warming, and the carbon balance of boreal forests, Ecol. Appl., 5(2): 437-451.
- Kurz, W. A., M. J. Apps, B. J. Stocks, and W. J. A. Volney, 1994, Global climate change: disturbance regimes and biospheric feedbacks of temperate and boreal forests: Biotic Feedbacks in the Global Climate System: Will the Warming Speed the Warming?: Oxford Univ. Press, Oxford, UK.
- Matson, M., and J. Dozier, 1981, Identification of subresolution high temperature sources using a thermal IR sensor, Phot. Eng. Remote Sensing, 47(9): 1311-1318.
- Scorer, R. S., 1987, Hot Spots and Plumes: Observation by Meteorological Satellite, Atmos. Environ., 21 (6): 1427-1435.
- Stocks, B. J., 1993, Global warming and forest fires in Canada, For. Chron., 69: 290-293.
- Stocks, B. J., D. R. Cahoon, Jr., W. R. Cofer III, and J. S. Levine, 1996, Monitoring large scale fire behavior in northeastern Siberia using NOAA-AVHRR satellite imagery, Biomass Burning and Global Change, MIT Press, Cambridge, MA. (in press)
- Whittaker, R. H., and G. E. Likens, 1975, Primary Production: The biosphere and man: Primary Productivity of the biosphere, Ecological Studies 14, Springer Verlag, New York.
- Wotton, B. M., and M. D. Flannigan, 1993, Length of fire season in a changing climate, For. Chron., 69: 187-192.

### BIOGRAPICAL SKETCH

Donald R. Cahoon, Jr. is a research scientist with the National Aeronautics and Space Administration, Langley Research Center, Hampton, Virginia. He has worked as an analyst of satellite, aircraft, and meteorological data sets for over 11 years. During the last 7 years he has specialized in satellite remote sensing. His primary applications have been forest fire monitoring and surface radiation budget studies.

Brian J. Stocks is a senior forest fire scientist with the Canadian Forest Service in Sault Ste. Marie, Ontario. Over the past 28 years he has specialized in the development of forest fire danger rating and fire behavior prediction systems, and is currently involved in the investigation of cause-and-effect relationships between global change and forest fires.

# Large Area Forest Cover Assessment: Effects of Misregistration in a Double Sampling Approach with Coarse and High Resolution Satellite Images

Christoph Kleinn<sup>1</sup>, Berthold Traub<sup>1</sup>, Matthias Dees<sup>1</sup>

**Abstract** When combining satellite data of considerable different spatial resolution the effect of misregistration is one of the technical issues to be addressed. This simulation study is based upon subcontract research carried out in the framework of the TREES project (Tropical Ecosystem Environment Observation by Satellites, Joint Research Centre, Ispra, Italy). Overall objective is a global tropical forest cover estimate and the production of a forest map. In a first phase of the inventory approach a complete coverage by the coarsely resolving NOAA AVHRR satellite images is provided. To improve the forest area estimates derived from this image set a sample of Landsat-TM scenes was selected in a second phase: In corresponding frames/blocks of the same geographical area forest cover percent<sup>2</sup> in AVHRR and TM was recorded. These pairs of values formed the input variables for a calibration regression.

To assume perfect geographic registration is certainly not realistic. Some effects of misregistration between the coarse and high resolution image frames/pixel blocks on the resulting regression are studied in this paper in the form of a simulation study. The effects can be considerable, particularly when small blocks of coarse resolution pixels are to be registered.

## 1. INTRODUCTION AND OBJECTIVES

Global monitoring of forest cover is a current issue in the context of the destruction of tropical forests and in the discussion about global climatic changes. Forest cover monitoring systems are to provide sound information on state and changes in forest cover. Satellite remote sensing plays an important role there: NOAA AVHRR has proved to have favourable characteristics to discriminate vegetation from other land cover and is frequently used in environmental monitoring (Ehrlich et al. 1994). LANDSAT TM has in the 1980's developed to be the standard high resolution sensor in many forestry applications. These two systems have quite different characteristics with respect to temporal, spatial and spectral resolution, in detail described in standard textbooks.

The high temporal frequency and low cost of the coarsely resolving AVHRR (about 1 km x 1 km) is well suited to provide in a first phase a more or less complete coverage of the region of interest. This results in an equally 'coarse' estimation of forest cover. The estimate might then in a second phase be improved through a sample of the much higher resolving LANDSAT TM, which is much more detailed but also much more expensive. Regression technique is used to

<sup>1</sup> Abteilung für Forstliche Biometrie, Universität Freiburg, Werderring 6, D-79085 Freiburg, Germany (kleinn@forst.uni-freiburg.de)

<sup>2</sup> The acronym *fcp* is used throughout the paper for *forest cover percent*

predict 'true' forest cover from the coarse information delivered by AVHRR. 'True' forest cover is assumed to be represented by the TM forest cover. In terms of sampling theory this is a double sampling for regression approach (Cochran 1977). From an image interpretation and classification point of view the resulting regression can be regarded as calibration function.

To calculate the regression it is necessary to obtain geographically matching pairs of map frames, one of the coarse resolution map and one of the high resolution map. This procedure is also called *coregistration*. Two approaches how to calculate the regression are discussed in the literature:

(1) *Coregistration of forest percentages* from both data sources; an example is found in Nelson (1989). This approach is also used in a global tropical forest assessment in the TREES project, a general description of which can be found in Malingreau (1993). Units to be coregistered can be areas corresponding to the size of one *single* coarse resolution pixel, with only two values possible: 0 and 1, or *blocks of pixels*.

(2) *Coregistration of forest percentages in the high resolution image with spectral values of coarse resolution image*. Examples are found in Iverson et al. (1989), Päivinen and Pitkänen (1992) and Zhu and Evans (1992). In this approach the units to be coregistered are areas of the size of one single coarse resolution pixel.

Like all types of measurements coregistration is subject to error, too: The problem of *misregistration* arises. This paper investigates in the form of a simulation study possible effects of misregistration onto the calibration procedure. It is an extension of a former study presented in Kleinn et al. (1995). Approach (1) is pursued here, using blocks of pixels as registration unit.

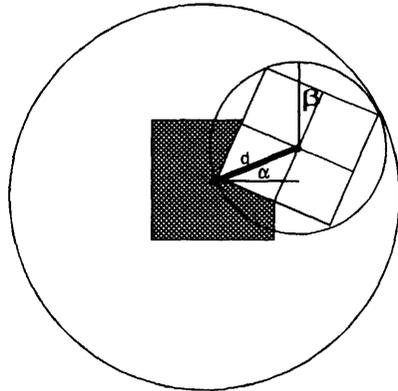
## 2. GENERAL DESCRIPTION OF MISREGISTRATION

Not many publications yet deal with the impact of misregistration in detail. One of the few articles focuses on change detection: Townshend et al. (1992) state that change assessments using satellite images of two different taking dates are affected by the level of registration, depending, of course, on several factors. They finally state that "high levels of registration must be achieved by operational monitoring systems if there is to be reliable monitoring of global change".

Assuming perfect classification procedures, the two data sources under consideration here (coarse and high resolution) would yield identical forest cover estimates. A simple linear regression between the data pairs would result in the one to one line. In reality, there are more or less significant deviations: Sensors of different spectral and spatial resolution 'see' things in a different way. Additionally geometric inaccuracies lead to *misregistration*: The block of high resolution pixels is not exactly matching the corresponding block of coarse resolution pixels.

Perfect registration would mean that (1) the centres of the two blocks match and that (2) their shape and (3) their alignment are the same. In Figure 1 these factors are depicted schematically. Misregistration means that not only the high resolution pixels in the true matching block have a chance to be registered, but also the pixels around them. Would the amount and type of misregistration be known, then one could calculate the probability of a pixel to be included in the registration process of one specific block. For perfect coregistration this probability would be 1

for the pixels in the matching block, 0 elsewhere. In the presence of misregistration the probability is high in the matching block but greater than 0 in a certain area around it. It is 0 only beyond the maximum misregistration distance. Under the assumption that misregistration is an isotrope process this area is drawn in Figure 1 as a circular envelope.



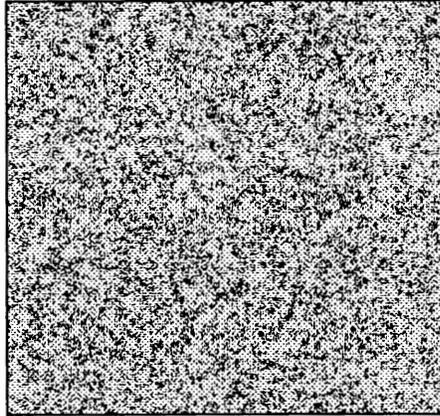
**Figure 1:** *Large circle:* Area in which the registration probability of the pixels of the high resolution images is greater than 0. *Small circle:* Area of pixels having an inclusion probability greater 0 for a 'rotating' displaced block with fixed distance  $d$  and shift angle  $\alpha$ .

Assumed that the actual registration of individual sample blocks in two different images is a random process, misregistration leads to a bias when estimating fcp: The expectation of the attribute (forest cover) measured in the high resolution image is *not* the value of the true matching block. This adds another source of variability (error) into the target regression between forest cover estimates of high and coarse resolution pixel blocks.

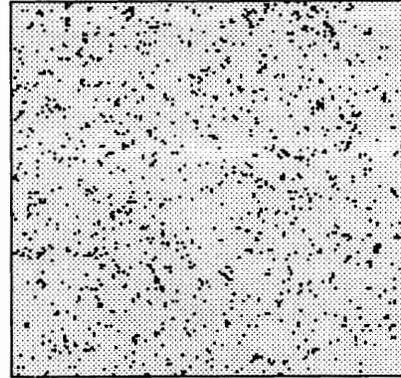
### 3. MATERIAL

Computer generated forest/non-forest (1/0) maps of size 9000x9000 pixels were used to investigate the effects of misregistration. A set of 'homogeneous' maps was created by randomly locating clusters of dots. These dots have diameters 3, 30 and 100 pixels for the map sets 1a-1e, 2a-2e and 3a-3e, respectively. The total fcp goes from 10% (in maps 1a, 2a, 3a) over 30%, 50%, 70% to 90% (in maps 1e-3e). Three inhomogeneous maps (maps 4 to 6) consist of regions with different structures. Examples of these maps at original, high resolution are given in Figure 2, left hand side.

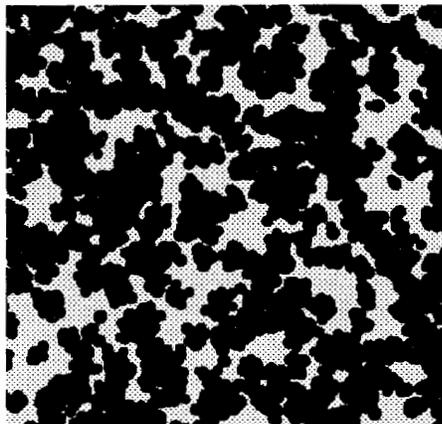
The high resolution maps were gradually degraded to produce coarse resolution maps. Square pixel blocks of size  $n \times n$  pixels of the original image were collapsed to one new coarse resolution pixel, with  $n$  taking on the values 10, 25 and 50. Depending on the forest cover within the  $n \times n$  high resolution pixels the attribute forest or non-forest was assigned to the new coarse resolution pixel. If fcp exceeded 30%, the attribute 'forest' was assigned. In Figure 2, right hand side, results of the degradation are shown. For the degraded maps only the 'core-region' of 8000x8000 pixels is shown in Figure 2, the region to which the analysis is limited. This allows for a buffer frame surrounding the analysed region, thus facilitating the treatment of edge effects in the simulations.



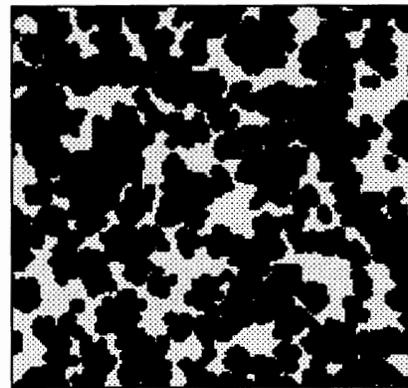
Map 1a (10% cover): Original resolution



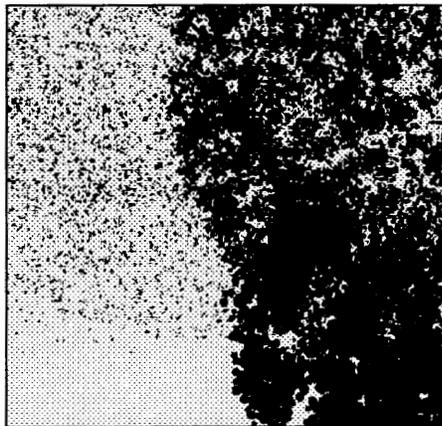
Map 1a: Degraded (level 50, see text)



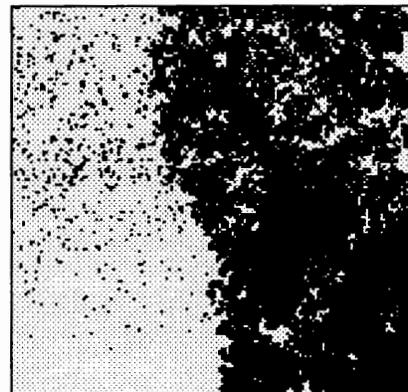
Map 3d (70% cover): Original resolution



Map 3d: Degraded (level 50, see text)



Map 4 (44.4% cover): Original resolution



Map 4: Degraded (level 50, see text)

**Figure 2: Sample maps as used in the simulation study. Left hand side: Original maps of size 9000x9000 pixels. For the degraded maps (right hand side) only the core-region corresponding to 8000x8000 pixels in original resolution is depicted.**

**Table 1: 'Forest' cover percent in the original and degraded maps used in this study. The fcp values of the maps shown in Figure 2 are framed.**

Map	Original resolution	Degraded		
		10 x 10	25 x 25	50 x 50
<i>forest cover percent of the total map</i>				
Map 1a	10.0	13.6	12.9	5.4
1b	30.0	35.2	40.1	40.1
1c	50.0	63.1	77.3	88.8
1d	70.0	83.8	95.5	99.5
1e	90.0	97.8	99.9	100.0
Map 2a	10.0	8.5	9.0	9.8
2b	30.0	32.0	33.9	36.5
2c	50.0	52.4	54.5	57.0
2d	70.0	72.1	74.7	78.3
2e	90.0	91.2	93.1	95.2
Map 3a	10.0	10.4	10.7	11.0
3b	30.0	30.2	30.8	31.6
3c	50.0	51.0	51.6	53.1
3d	70.0	70.6	71.5	72.9
3e	90.0	90.6	91.1	92.0
Map 4	44.4	48.1	50.1	51.5
Map 5	32.6	45.3	51.7	57.5
Map 6	34.1	35.4	37.8	40.0

The image degradation technique leads to a change in total forest cover as illustrated in Table 1. It is a very simple technique, which probably does not mirror very realistically the properties of two real map sets. But it was felt that this was sufficiently realistic to investigate the general properties of misregistration. With real data sets the differences in fcp values would be due to differences in spatial resolution, taking date, atmospheric conditions etc. and, of course, differences in image interpretation and classification.

#### 4. METHODS

When modeling misregistration one has to make assumptions on the distance and/or direction *distribution* of the deviations relative to the true location. The registration process is modelled as purely random process here. The factors which were included in this simulation and which were subject to variation at several levels are listed in Table 2. Non-matching shapes of the two blocks to be are not taken into respect in this study. For each combination of the listed factor levels a systematic grid of coarse resolution sample blocks was superimposed onto the original, high resolution maps. For all blocks at first the fcp of the true matching high resolution block was determined. Then the high resolution block was displaced 100 times according to the probability assumptions given in Table 2. For each block, mean and standard deviation of the fcp values resulting from the 100 replications were recorded.

#### 5. RESULTS

***Statistics of the differences between truly matching pixel block and the mean (expectation) of the 100 simulated misregistered pixel blocks:*** In Figure 1 it was illustrated that a bias is suspected when estimating the fcp as an expectation of

misregistered pixel blocks. In some results are given for the maps shown in Figure 2, for degradation level 50. The bias (*mean of the differences* for all blocks analysed throughout the whole map) is negligible for all maps investigated.

**Table 2: List of factors included in the model study**

<i>Factor</i>	<i>Description and factor levels</i>
Area structure	Artificially generated maps: Some produced with a homogeneous generation process, some with several processes overlaid (heterogeneous)
Total forest cover	Total forest cover is 10%, 30%, 50%, 70% and 90% for the homogeneous maps and 32% to 44% for the heterogeneous ones
Spatial resolution of coarse resolution image	Side length of the square pixels of the coarse resolution maps correspond to 10, 25 and 50 pixels of the high resolution image
Forest/non-forest rule to be used for image degradation	Minimum 'crown' cover percent in the degraded image for a pixel to be assigned to the class 'forest' is fixed to be 30%
Block size	Coarse resolution square pixel blocks of 2, 5, 10, 20 pixels side length
Parameters of misregistration	Misregistration described in Figure 1. Maximum distance is 4 pixels of coarse resolution following a linearly decreasing pdf (mean distance = $4/3 = 1.333$ pixels, standard deviation = $\sqrt{16/18} = 0.943$ pixels). Isotropy is assumed. The coregistered blocks may rotate up to $\pm 10$ degrees, following a triangular probability density function symmetric around 0 degrees.

For the single sample blocks - i.e. not using the mean of the 100 replications per block - the difference between truly matching pixel block and misregistered pixel block can be considerable. This can be seen with the range and standard deviation of the differences as given in Table 3: For Map 3d and block size 2 x 2 the standard deviation is about 10% fcp, having a range of differences from -41% to +38%, meaning that one has to be aware of large variation when making blockwise evaluations. For a block size of 10 x 10 or even 20 x 20 this variation is much less.

This variability of the differences is also a function of total forest cover, which can be observed when analysing for example the sequence of maps 1a to 1e. These maps were generated with the same algorithm of random clustering of dots of a diameter of 3 pixels. The difference is that Map 1a was only filled up to 10%, Map 1b up to 30% and so on. Map 1e has a 90% cover. In Figure 3 the standard deviation of the differences is shown over total forest cover: The highest variability is thus reached with the intermediate fcp of around 50%. With a forest cover of 10% or 90% variability is much less. This relationship is a quite obvious one when analysing the extreme cases: In the complete presence (or absence) of forest, misregistration has no effect at all.

**Calibration regression:** The regression is to predict blockwise the fcp estimate of the coarse resolution pixel blocks. Here, simple linear regression is used. Though it would be justified for practical reasons the intercept was not suppressed. Given the situation illustrated in Table 1 it is clear that one gets a *slope coefficient* different from 1 (in most cases smaller than 1), when predicting fcp of the high resolution map (dependent variable) using the fcp of the coarse resolution map as independent variable. This holds even for the situation of perfect coregistration.

In analysing the regression results we are interested in two relationships: **Regression 1** under perfect coregistration and **Regression 2** in the presence of

misregistration. In Table 4 general characteristics of the regressions are listed for the three sample maps addressed throughout this paper. Given the same map and the same degradation level there are considerable differences between the coefficients of Regression 1 and Regression 2 for the small block size 2, but the differences level out quite markedly for the largest block size 20. Intercept and slope coefficients of the Regressions 1 and 2 are quite close for this large block size. As Regression 1 is the 'true' regression under perfect coregistration one sees that small block sizes lead to an incorrect calibration function and that only for very large block sizes Regression 2 approaches the coefficients of the true Regression 1.

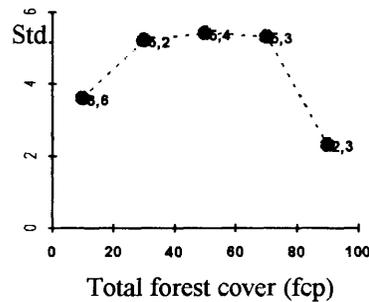


Figure 3: Standard deviation of differences between fcp of truly matching pixel block and the mean of the 100 misregistered pixel blocks over fcp

Table 3: Descriptive statistics of the differences between the fcp of the true matching high resolution block and the fcp results of the 100 simulated misregistered blocks (degradation level 50):

	Block size	Mean	Std Dev	Minimum	Maximum
Map 1a	2	0.000246	0.0356	-0.0931	0.1679
	5	0.000281	0.0103	-0.0326	0.0442
	10	0.000255	0.0036	-0.0098	0.0096
	20	0.000256	0.0011	-0.0029	0.0031
Map 3d	2	0.003002	0.1017	-0.4104	0.3791
	5	0.002945	0.0429	-0.1661	0.1418
	10	0.002647	0.0156	-0.0417	0.0568
	20	0.003288	0.0061	-0.0149	0.0177
Map 4	2	0.001279	0.0821	-0.4125	0.4526
	5	0.001563	0.0195	-0.1277	0.0690
	10	0.001645	0.0080	-0.0338	0.0274
	20	0.000914	0.0038	-0.0111	0.0125

Differences between the coefficients of Regression 1 and Regression 2 are most clear for map set 1, consisting of very small forest 'patches' compared to the spatial resolution of the degraded images. This holds - as mentioned in the preceding paragraph - particularly for small block sizes and low total forest cover. For larger patches and higher total fcp the effect is much smaller, though still there.

**Table 4: Basic statistics of the simple linear calibration regression for three sample maps (rmse=root mean square error, r2=coefficient of determination).**

**Map 1a (10% cover in original map)**

Deg lev.	Bl. size	Regression 1				Regression 2				within block std
		rmse	r2	Interc.	Slope	rmse	r2	Interc.	Slope	
10	2	0.0509	0.9177	0.0093	0.7028	0.0569	0.7789	0.0443	0.4413	0.1058
	5	0.0201	0.9618	0.0042	0.7404	0.0246	0.9283	0.0164	0.6483	0.0406
	10	0.0099	0.9798	0.0023	0.7546	0.0123	0.9650	0.0086	0.7055	0.0207
	20	0.0049	0.9873	0.0014	0.7611	0.0057	0.9816	0.0037	0.7416	0.0081
25	2	0.0507	0.7581	0.0451	0.4640	0.0447	0.4970	0.0750	0.2296	0.0769
	5	0.0234	0.8497	0.0339	0.5511	0.0232	0.7929	0.0466	0.4503	0.0302
	10	0.0120	0.8892	0.0291	0.588	0.0121	0.8711	0.0344	0.5445	0.0118
	20	0.0057	0.9091	0.0286	0.5924	0.0058	0.8980	0.0315	0.5675	0.0047
50	2	0.0507	0.4728	0.0848	0.3735	0.0338	0.3682	0.0938	0.2006	0.0531
	5	0.0238	0.5631	0.0792	0.4768	0.0200	0.5360	0.0842	0.3790	0.0195
	10	0.0113	0.6388	0.0777	0.5051	0.0107	0.6099	0.0805	0.4486	0.0078
	20	0.0061	0.7116	0.0761	0.5348	0.0059	0.7062	0.0770	0.5137	0.0026

**Map 3d (70% fcp in original map)**

Deg lev.	Bl. size	Regression 1				Regression 2				within block std
		rmse	r2	Interc.	Slope	rmse	r2	Interc.	Slope	
10	2	0.0331	0.9945	-0.0042	0.9978	0.0538	0.9845	0.0192	0.9592	0.0485
	5	0.0165	0.9985	-0.0078	1.0030	0.0229	0.9971	0.0020	0.9854	0.0343
	10	0.0101	0.9994	-0.0085	1.0040	0.0147	0.9986	-0.0014	0.9905	0.0345
	20	0.0059	0.9997	-0.0088	1.0044	0.0064	0.9997	-0.0064	0.9976	0.0235
25	2	0.0513	0.9857	-0.0107	0.9949	0.0757	0.9641	0.0445	0.9143	0.0972
	5	0.0247	0.9960	-0.0192	1.0067	0.0346	0.9915	0.0087	0.9642	0.0749
	10	0.0139	0.9983	-0.0210	1.0093	0.0188	0.9967	-0.0076	0.9875	0.0525
	20	0.0075	0.9991	-0.0197	1.0074	0.0089	0.9987	-0.0135	0.9954	0.0299
50	2	0.0690	0.9707	-0.0197	0.9875	0.1006	0.9168	0.0923	0.8297	0.1603
	5	0.0325	0.9905	-0.0375	1.0119	0.0444	0.9792	0.0198	0.9293	0.1079
	10	0.0170	0.9953	-0.0382	1.0128	0.0205	0.9926	-0.0116	0.9728	0.0581
	20	0.0090	0.9966	-0.0410	1.0168	0.0108	0.9950	-0.0347	1.0035	0.0267

**Map 4 (44% fcp in original map)**

Deg lev.	Bl. size	Regression 1				Regression 2				within block std
		rmse	r2	Interc.	Slope	rmse	r2	Interc.	Slope	
10	2	0.0625	0.9817	-0.0067	0.9734	0.0938	0.9540	0.0227	0.9080	0.0915
	5	0.0324	0.9943	-0.0147	0.9899	0.0402	0.9909	-0.0056	0.9677	0.0523
	10	0.0210	0.9973	-0.0161	0.9930	0.0232	0.9966	-0.0137	0.9847	0.0294
	20	0.0150	0.9985	-0.0165	0.9938	0.0154	0.9984	-0.0161	0.9898	0.0141
25	2	0.0870	0.9592	-0.0061	0.9336	0.1117	0.9226	0.0325	0.8535	0.1167
	5	0.0461	0.9865	-0.0187	0.9588	0.0497	0.9837	-0.0109	0.9413	0.0494
	10	0.0329	0.9924	-0.0201	0.9616	0.0338	0.9919	-0.0180	0.9538	0.0273
	20	0.0274	0.9943	-0.0198	0.9610	0.0273	0.9943	-0.0198	0.9574	0.0120
50	2	0.0983	0.9406	0.0127	0.8716	0.1135	0.9109	0.0441	0.8081	0.1048
	5	0.0598	0.9750	0.0001	0.8959	0.0598	0.9742	0.0049	0.8836	0.0427
	10	0.0482	0.9825	-0.0001	0.8962	0.0477	0.9826	0.0013	0.8904	0.0235
	20	0.0411	0.9868	0.0017	0.8927	0.0409	0.9868	0.0027	0.8893	0.0128

For map 3d, consisting of large forest patches (diameter 100 original pixels each) and having a total forest cover 70% this approximation is quite good even for the small block sizes. For degradation level 10 for example the slope coefficient is 0.9592 for block size 2 ('true' coefficient: 0.9978) and 0.9976 for block size 20, being almost equal to the 'true' 1.0044. For Map 1a, consisting of very small patches with a total cover of only about 10%, there is a considerable difference between the slope coefficient for block size 2 (0.4413) and the 'true' value of 0.7028, while for block size 20 the difference between 'misregistered' slope (0.7416) and 'true' slope (0.7611) is small.

The differences in slope, however, do not lead to an error of estimation of overall fcp. This is illustrated in Table 3. The differences are leveled out by the adverse differences in intercept: Lower slope coefficients go together with higher intercepts. In Map 1a the differences in intercept are relatively small. But one has to consider that the vast majority of observations in this map is in the class fcp=0, as total forest cover is only 10%. So small differences in intercept have a large effect.

When the calibration regressions based upon misregistered blocks are used for map production (calibrating the fcp of each coarse resolution pixel block) a clear error is introduced. For larger fcp values the true value will be underestimated, for very small values overestimated. Negative intercepts as they occur in Table 4 are normally not acceptable in a calibration regression: It would mean that a forest-free coarse resolution pixel block is calibrated from an fcp of 0.0 to a negative fcp. In our simulation study, however, the calibration regression was intentionally not forced to be non-negative.

The root mean square error (rmse) is in most cases higher for Regression 2 what would be expected as misregistration introduces more variability. The rmse, however, is not an adequate measure here as calculations were made using per block the *mean* of the fcp of 100 simulations. Thus, if one is interested in the variability for one single set of misregistered blocks, one would have to take additionally into account the variability within the individual blocks.

## 7. CONCLUSION

Three major factors can be identified which determine the magnitude of the misregistration effect as to this simulation study:

- *Magnitude of misregistration.* The bigger the dislocation distance the bigger the difference in fcp values registered. This obvious relation, however, must also be seen in interacting context with the two other factors.

- *Size of pixel blocks* to be registered. The bigger the block size the less the effect of misregistration. For small pixel blocks the effect can be considerable. This relation interacts very much with the third important factor, the

- *spatial structure and fcp* of the region of interest. For very high and very low fcp the misregistration effect is smaller than for intermediate fcp's. If the forest patches are large in comparison to the size of the pixel blocks to be coregistered, the effect of misregistration on fcp estimation is smaller.

The bias of the total forest cover estimate that was suspected on the basis of theoretical considerations could not be confirmed to be significant in this

simulation study. However, the effect of geometric misregistration can considerably add to the *error* of the estimation of total fcp.

The effect on *map accuracy* appears more important. Though the overall forest cover estimate was found to be almost not affected at all, there was a considerable effect on the coefficients of the calibration regression. This deviation of the 'misregistered' calibration regression from the true one leads to leads in the cases investigated here to an underestimation of high fcp values and to an overestimation of low ones.

It may be concluded that misregistration is one source of error to be considered when simultaneously analysing images of different resolution under a double sampling approach.

**ACKNOWLEDGEMENT:** The authors thank Mr. Olaf Kügler, Mr. Jürgen Harm and Mrs. Manuela Bacher for their help during the computational procedures.

## REFERENCES

- Cochran, W.G. (1977): Sampling Techniques. 3rd edition. John Wiley&Sons.
- Ehrlich, D., J. E. Estes and A. Singh (1994): Applications of NOAA-AVHRR 1km data for environmental monitoring. *Int. Journal of Remote Sensing* 15(1): 145-161.
- Iverson, L. R.; Cook, E. A. and Graham, R. L. (1989), Estimating Forest Cover over Southeastern United States using TM-calibrated AVHRR Data. In *Proceedings of Global Natural Research Monitoring and Assessments: Preparing for the 21st Century*, Venice, Italy, September, VOL 3, pp. 1252-1262.
- Justice, C. O., Markham, B. L., Townshend, J. R. G., and Kennard, R. L. (1989), Spatial degradation of satellite data. *Int. Journal of Remote Sensing* 10(9): 1539-1561.
- Kleinn, C., B. Traub, M. Dees (1995): Calibrating Forest Cover Estimates from Coarse Resolution Images with Samples of High Resolution Images - A Model Study on Possible Effects of Misregistration -. Paper presented at IUFRO World Congress 1995, Tampere, Finland.
- Malingreau, J.-P. (1993): Introduction to the TREES Project: Objectives, Features, Achievements and Plans. TREES Validation Workshop, Joint Research Centre, Ispra, Italy, March 9-10, 1993.
- Nelson, R. (1989), Regression and Ratio Estimators to Integrate AVHRR and MSS Data, *Remote Sens. Environ.* 30:201-216.
- Päivinen, R., and Pitkänen, J. (1992), Calibrating AVHRR Data with Landsat TM Data for Tropical Closed Forest Assessment in Ghana, in *Remote Sensing and Permanent Plot Techniques for World Forest Monitoring*, IUFRO S.4.02.05 International Workshop Remote Sensing and World Forest Monitoring, edited by Lund, H. G., Päivinen, R., and Thammincha, S., Pattaya, Thailand pp. 95-102.
- Townshend, J. R. G., Justice, C. O., Gurney, C. McManus, J. (1992), The impact of misregistration on change detection. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 30, No. 5: 1054-1060.
- Zhu, Z., and Evans, D. L.(1992), Mapping Midsouth Forest Distributions -AVHRR satellite data and GIS help meet RPA mandate. *Journal of Forestry* Vol. 90, No.12, December 1992, pp.27-30.

# Linear Mixture Modeling with Autocorrelated Errors

Jayantha Ediriwickrema<sup>1</sup>, Siamak Khorram<sup>2</sup>, Marcia Gumpertz<sup>3</sup> and John Brockhaus<sup>4</sup>

---

**Abstract.**—The linear mixture model assumes that model errors are spatially uncorrelated. Spatial continuities exist in most geographical data. The data values close to each other are likely to have similar spectral characteristics. Especially in remotely sensed data, spatial autocorrelations are present among pixels. Residuals of the Advance Very High Resolution Radiometer data are examined for any deviation from the random assumption. The correlation among the residuals of the neighboring pixels and Moran's I statistic demonstrate the significance of non-random distribution of residuals. This study develops an autocorrelated error model to calibrate the linear mixture model. Seven different autocorrelation patterns are considered, and for each pattern, the linear mixture model is calibrated and land use/land cover class fractions are estimated from the Advance Very High Resolution Radiometer data.

---

## INTRODUCTION

The effects of spatial autocorrelation are not clearly understood in unmixing mixed pixels by the linear mixture model (LMM). Autocorrelated errors are common in remotely sensed data (Campbell and Kiiveri 1988). Campbell and Kiiveri (1988) described point spread function of the sensor and atmospheric effects as causes for the spatial autocorrelation among neighboring pixels. The radiance values of the neighboring pixels may show similar spectral characteristics due to scattering and similar land use/land cover (LU/LC) class patterns. These effects become apparent with increasing coarseness of the spatial resolution. The autocorrelated structure of remotely sensed data is widely used as additional knowledge in image classification. Belward (1992) used spatial attributes: contextual information and autocorrelated errors to evaluate the Advance Very High Resolution Radiometer (AVHRR) data in environmental monitoring. Campbell and Kiiveri (1988) discussed the advantages of using the spatial relations of the neighboring pixels in image classification, but they did not observe a significant advantage due to the spatial autocorrelation when the spectral class separations were high. Furthermore, no significant improvement was noticed in overall accuracy when spatial autocorrelations were included in the classification process.

In the empirical method of unmixing mixed pixels by LMM, calibration coefficients are estimated by multivariate regression analysis. The calibration coefficients are usually estimated assuming the model errors are spatially uncorrelated with constant variance. Iverson et al. (1989) used multivariate regression analysis to develop an empirical relationship between the AVHRR data

---

<sup>1</sup> Graduate Assistant, Computer Graphics Center, NCSU, Raleigh, NC.

<sup>2</sup> Professor and Director, Computer Graphics Center, NCSU, Raleigh, NC, and Dean of International Space University, Parc d'Innovation, Communaute Urbaine de Strasbourg, Blvd. Gonthier d' Andernach, 67400 Ilkirch, France

<sup>3</sup> Associate Professor, Department of Statistics, NCSU, Raleigh, NC.

<sup>4</sup> Assistant Professor, Mapping, Charting and Geodesy, Department of Geography and Environmental Engineering, United States Military Academy, West Point, NY.

and the forest cover. Pech et al. (1986) examined multivariate calibration methods to estimate vegetation cover. In all these studies, the model errors were assumed to be spatially uncorrelated with constant variance.

Multivariate regression models with autocorrelated errors are common in statistics. Nevertheless, the spatial autocorrelation effects are usually assumed insignificant in unmixed mixed AVHRR pixels. Autocorrelated error models are widely used for spatially correlated data accounting for the interaction between neighboring locations. Upton and Fingleton (1985) described a generalized least squares (GLS) regression method for spatially autocorrelated models. Basically, models with normally distributed spatially autoregressive errors fall in two schemes: simultaneous and conditional autoregressive schemes. The autocorrelation parameters in this study were estimated by a maximum likelihood method described by Upton and Fingleton (1985) based on a simultaneous autoregressive scheme. Seven spatial autocorrelation patterns were compared with the model that assumed errors were spatially uncorrelated.

## METHOD

The study area, North Carolina Piedmont showed a wide variation of LU/LC patterns. The diversity of the LU/LC class patterns ranged from developed urban to various distributions of residential, agriculture, and forests. The frequency of the local variation was high within this area. A significant amount of autocorrelation was anticipated among the AVHRR pixels especially due to mixed pixel effects.

### Data

Classified Landsat Thematic Mapper (TM) data and NOAA AVHRR data corresponding to the North Carolina Piedmont data set used by Khorram et al. (1994) were used in this study. Two sets of ASCII data files were developed in this process. One set was developed from the classified Landsat TM data and the other set was derived from the AVHRR data. The file created from the Landsat TM data described LU/LC class fractions within each one square km area corresponding to the AVHRR pixels. The ASCII files created from the AVHRR data contained digital number values in 14 spectral bands (Ediriwickrema 1995).

### The LMM with Autocorrelated Errors

The LMM assumed that there were no interactions between the six LU/LC classes. It also assumed that the resultant radiation of each pixel could vary only due to random noise. Correlation between AVHRR pixel values and atmospheric errors were also considered as insignificant. Under these assumptions, composite radiation of mixed pixels was related to the component LU/LC class fractions and to the spectral values of the LU/LC class pure pixels by Eq. 1. A pure pixel was defined as a pixel that was totally covered by only one defined LU/LC class.

$$A_{jk} = \sum_{i=1}^c F_{ji} R_{ik} + \alpha_{jk}, \quad [1]$$

where  $A_{jk}$  = pixel value of the  $j^{\text{th}}$  AVHRR pixel in band  $k$ ,  $F_{ji}$  = fraction of the  $i^{\text{th}}$  LU/LC class within the  $j^{\text{th}}$  AVHRR pixel,  $R_{ik}$  = calibration coefficients of the  $i^{\text{th}}$  LU/LC class in band  $k$ ,  $\alpha_{jk}$  = random noise of the  $j^{\text{th}}$  AVHRR pixel in the band  $k$ , and  $c$  = the number of LU/LC classes. The error term “ $\alpha_{jk}$ ” was assumed to be

spatially correlated. The Gauss-Markov condition in the OLS regression is not satisfied when the residuals are spatially correlated. One model for spatially correlated errors is the simultaneous autoregressive error model. In this model the error for one pixel depends upon the errors of neighboring pixels. The error vector  $\alpha$  for one band having “n” pixels is  $\alpha = \rho W\alpha + u$  (Upton and Fingleton 1985), which gives  $u_{nx1} = (I_{n \times n} - \rho W_{n \times n})\alpha_{nx1}$  and  $\rho$  is a constant -- autocorrelation parameter.  $W$  is a proximity matrix in which each row indicates which pixels are neighbors in the autocorrelation scheme,  $u$  = error vector that satisfies the Gauss-Markov conditions, and  $I$  = identity matrix. Since the elements of “u” are uncorrelated, the simultaneous autoregressive model can be fitted by OLS after pre multiplying both sides by  $(I - \rho W)$ .

$$(I_{n \times n} - \rho W_{n \times n})A_{n \times b} = (I_{n \times n} - \rho W_{n \times n})F_{n \times c}R_{c \times b} + u_{n \times b} \quad [2]$$

In other words,  $A^*_{n \times b} = F^*_{n \times c}R_{c \times b} + u_{n \times b}$ , where  $A^*_{n \times b} = (I_{n \times n} - \rho W_{n \times n})A_{n \times b}$  and  $F^*_{n \times c} = (I_{n \times n} - \rho W_{n \times n})F_{n \times c}$ . Then the estimates of the calibration coefficients ( $\hat{R}_{c \times b}$ ) are written as follows.

$$\hat{R}_{c \times b} = (F^{*T}_{c \times n} F^*_{n \times c})^{-1}_{c \times c} F^{*T}_{c \times n} A^*_{n \times b} \quad [3]$$

A reasonable  $\rho$  value for each spectral band was required to estimate  $R_{c \times b}$ . In the simultaneous autoregression,  $\rho$  was chosen such that it maximized the log likelihood function “ln(L)” in Eq. 4.

$$\ln(L) = -\frac{n}{2} \ln \hat{\sigma}^2 + \ln |I - \rho W|, \quad [4]$$

where  $\hat{\sigma}^2 = \hat{u}^T \hat{u} / n$  and  $\hat{u} = (I_{n \times n} - \rho W_{n \times n})A_{n \times b} - (I_{n \times n} - \rho W_{n \times n})\hat{A}_{n \times b}$  (Upton and Fingleton 1985). In Eq. 4 the determinant  $|I - \rho W|$  is not easy to compute for large data sets. Upton and Fingleton (1985) describe formulas to avoid computational complications as well as a way to compute  $|I - \rho W|$  for large data sets. These formulae are applicable for only one type of proximity matrix, the Rook’s case (Upton and Fingleton 1985). Therefore, a 20 column and 20 row subset was selected from the upper half by considering available system resources to reduce computational burden. This subset was used to calibrate the LMM. A similar size subset was also selected from the lower half of the image to estimate LU/LC class fractions, and thereby examine the effects of autocorrelated errors in the LMM.

### Weightings

The proximity matrix ( $W$ ) is constructed according to the autocorrelation pattern. Upton and Fingleton (1985) described three weightings named Rook, Bishop and Queen. Rook weighting considers only adjacent pixels within a row or column to be neighbors. The Bishop weighting considered only the cells that touch at the corners to be neighbors, while the Queen model considers all surrounding cells, i.e., touching corners and sides to be neighbors. In this study, seven different weighting structures were considered (Ediriwickrema 1995).

### **Proximity matrix**

In the proximity matrix, the cells that account for contiguity are assigned one, and the remaining cells are assigned zero. Each row in the proximity is divided by its total so that row totals are all equal to one. For example, a lattice of 3 x 3 pixels results in a 9 x 9 proximity matrix. The proximity matrix for the weighting Rook's case will be similar to the one that is shown in (Ediriwickrema 1995). In that proximity matrix, the boundary pixels are weighted differently than the other pixels. Only the available surrounding pixels are accounted for; there are many other possibilities for weighting pixels on the edges. When all pixels are not weighted with the same number of pixels, the scaling results in an asymmetric "W" matrix.

### **Analysis of Spatial Continuity in the Model Residuals**

The spatial autocorrelation structure of the model residuals are commonly examined by semivariogram plots of the data, h-scatter plots (Isaaks and Srivastava 1989) of the errors, Moran's I statistic (Upton and Fingleton 1985). The residuals from the regression of AVHRR data on the LU/LC class fractions using ordinary least squares (OLS) were used to analyze the spatial continuity in the model. Ediriwickrema (1995) describes seven different spatial autocorrelation patterns. The spatial autocorrelation coefficient of the pixel residuals from an OLS fit of the LMM to their neighboring (as defined for each weighting) pixel residuals was calculated for each spectral band. The significance of spatial autocorrelation of the OLS residuals was tested using Moran's I statistic.

### **Autocorrelation parameters**

The magnitude and the sign of the autocorrelation parameters measure the scale and the direction of spatial continuity. The proximity matrices are commonly designed with row totals equal to one, which restrains  $\rho$  to be between -1 and 1. Haining (1990) pointed out that  $\rho$  is an eigenvalue of  $W^{-1}$ , so  $\rho$  must lie between  $1/\lambda_{\max}$  and  $1/\lambda_{\min}$ , where  $\lambda_{\max}$  is the largest eigenvalue of  $W$  and  $\lambda_{\min}$  is the smallest eigenvalue of  $W$ . As described in section "The LMM with Autocorrelated Errors," the  $\rho$  value for each band was calculated such that it maximized "ln(L)". In this study, the log likelihood function "ln(L)" values were evaluated for different  $\rho$  values ranging from -1 to 1 with an increment of 0.01. The  $\rho$  value corresponding to the maximum "ln(L)", was selected as the autocorrelation parameter estimate of the respective spectral band.

### **Calibration of the LMM**

Calibration coefficients were estimated by Eq. 5 based on spectral bands. For each spectral band, six coefficients were estimated representing all LU/LC classes.

$$\hat{R}_{cx1} = (F^*_{cxn} F^*_{nxc})^T_{cxc} F^*_{cxn} A^*_{nx1} \quad [5]$$

Each time when the calibration coefficients were estimated for a band, residual errors ( $\hat{u}_{nx1}$ ) were also calculated from the same data. Finally, the overall calibration coefficients matrix ( $\hat{R}_{cxb}$ ) and the residual error matrix ( $\hat{u}_{nxb}$ ) were developed. The variance-covariance matrix of the residual errors was calculated by

$\hat{\Sigma}^*_{bxb} = \hat{u}^T_{bxn} * \hat{u}_{nxb} / n$ . In order to compare the autocorrelated models to the model that assumed errors are spatially uncorrelated, calibration coefficients were separately estimated assuming the model errors were spatially uncorrelated.

### Estimation of the LU/LC Class Fractions from the AVHRR Data

The North Carolina Piedmont data set was divided into two portions. From the upper portion, the subset ( $A_{nxb}$ ) was selected to calibrate the LMM model. Subset ( $B_{mxb}$ ) was selected from the second portion to estimate the LU/LC fractions ( $G_{mxc}$ ) and assess the autocorrelated errors in the LMM.  $B_{mxb} = G_{mxc} Q_{cxb} + w_{mxb}$ , where  $B_{mxb}$  = digital values of “m” AVHRR pixels in “b” spectral bands,  $Q_{cxb}$  = calibration coefficients of “c” LU/LC classes in “b” spectral bands, and  $w_{mxb}$  = residual error matrix. Let  $Q_{cxb} = \hat{R}_{cxb}$  and assuming the measurement errors associated with  $\hat{R}_{cxb}$  are insignificant, the regression model can be written as  $B_{mxb} = G_{mxc} \hat{R}_{cxb} + w_{mxb}$  and can also be re-written by taking transpose as  $B^T_{bxb} = \hat{R}^T_{cxc} G^T_{cxm} + w^T_{bxm}$ .

We propose a simple model incorporating spatial correlation within and between spectral bands. Let  $b = \text{Vec}(B^T) = [b_1^T b_2^T b_3^T \dots b_{14}^T]^T$ , where  $b_k$  is an  $m \times 1$  vector of AVHRR values for the  $k^{\text{th}}$  spectral band. Let

$$\Sigma = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \dots & \sigma_{1,14} \\ \vdots & & & \\ \sigma_{14,1} & \sigma_{14,2} & \dots & \sigma_{14,14} \end{bmatrix} \quad [6]$$

denote the covariance matrix among bands within a pixel. Let  $C_k = (I - \rho_k W)^{-1}$  denote the square root matrix of the spatial autocorrelation for band  $k$ . Then let the covariance matrix for all bands and all pixels be

$$\text{Var}(b) = \begin{bmatrix} C_1 C_1^T \sigma_{1,1} & C_1 C_2^T \sigma_{1,2} & \dots & C_1 C_{14}^T \sigma_{1,14} \\ \vdots & & & \\ C_{14} C_1^T \sigma_{14,1} & C_{14} C_2^T \sigma_{14,2} & \dots & C_{14} C_{14}^T \sigma_{14,14} \end{bmatrix} \quad [7]$$

This covariance model says that the covariance among two bands in two different pixels is just the product of the components of spatial variance for each of the two bands and the covariance among bands “k” and “l” within a pixel  $\sigma_{kl}$ . The estimated GLS estimator of the LU/LC class fractions for the  $j^{\text{th}}$  pixel is then

$$\hat{g}_{j, \text{GLS}} = (\hat{R} \hat{V}_j^{-1} \hat{R}^T)^{-1} \hat{R} \hat{V}_j^{-1} b_j, \quad [8]$$

where  $\hat{V}_j$  is a  $14 \times 14$  matrix of estimates of  $\text{Var}(b)$  corresponding to the  $j^{\text{th}}$  pixel with the elements of  $\Sigma$  estimated by  $\hat{\Sigma}^*_{bxb} = \hat{u}^T_{bxn} * \hat{u}_{nxb} / n$ , and  $b_j$  is a  $14 \times 1$  vector of AVHRR values for the  $j^{\text{th}}$  pixel.

The LU/LC class fractions were found to be best estimated by the constraint least square regression method (Ediriwickrema 1995). The same method was used in this study to estimate the LU/LC class fractions from the AVHRR data. The LU/LC class fractions were estimated separately considering each autocorrelated pattern, and assuming errors were spatially uncorrelated (Ediriwickrema 1995).

## Validation

To examine the developments in the autocorrelated error model to the OLS model, the residual sum of squares was calculated for four sample areas. From the upper left of each quadrant, a 20 x 20 sample area was selected.

## RESULTS

### Correlation of Model Residuals

The correlation of the model residuals among the neighboring pixels for each weighting is described in Figure 1. Except for the bands 11 and 12 generally all other bands show significant correlation among the neighboring pixels in all seven spatial autocorrelation patterns. Horizontal and vertical spatial autocorrelation patterns resulted in high correlation related to the diagonal spatial patterns.

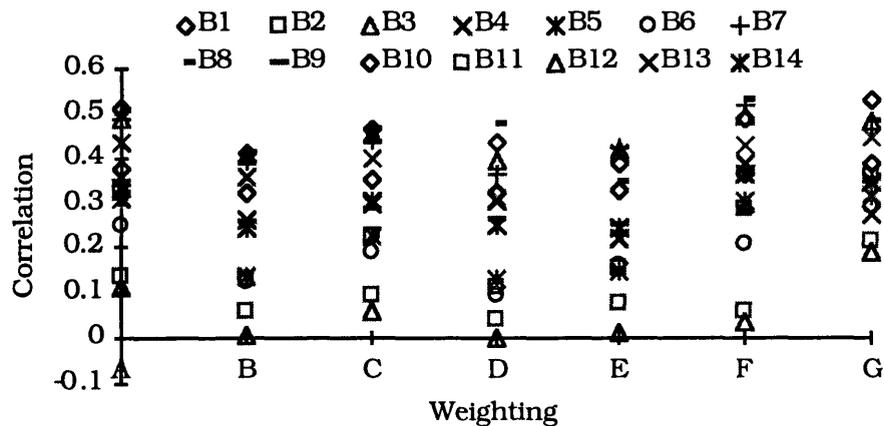


Figure 1.—Correlation of model residuals among neighboring pixels in each weighting and each spectral band.

### Significance of Autocorrelated Errors

Except for the z-statistics in bands 11 and 12 in correlation patterns “b”, “d”, “e” and “f” (Ediriwickrema 1995), all other z-statistics are greater than the upper one-tail critical value at the 95% confidence level (1.645) (see Figure 2). The z-statistic values in all other spectral bands in all autocorrelation patterns are clearly greater than 1.645.

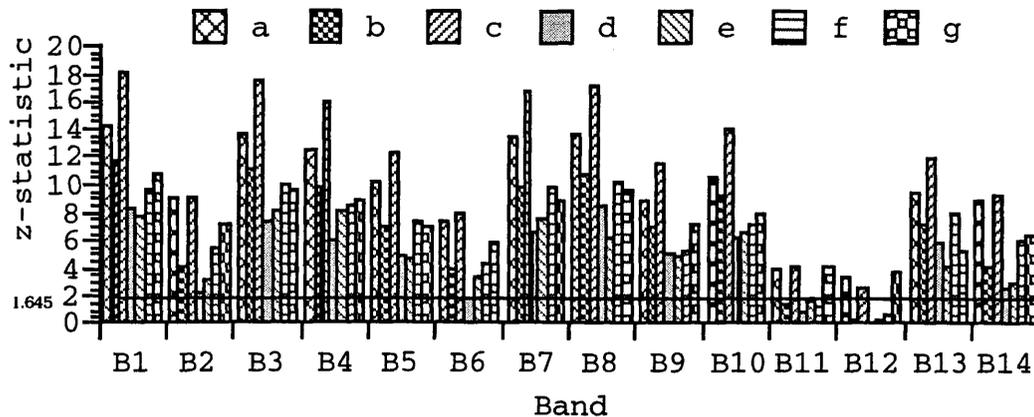


Figure 2.—z-statistics of each AVHRR band in each autocorrelated pattern.

### Validation of the Autocorrelated Error Model

The difference of the sum of squares of the residuals between the OLS and the autocorrelated error model are shown in Figure 3. The residual sum of squares was reduced in the autocorrelated error model for most of the bands in all four quadrants. The spectral bands acquired in late June and in mid July show remarkable improvement. Except for bands five and six in the upper right quadrant, all other bands in all quadrants show an improvement with the autocorrelated error model. The magnitude of the improvement is dependent on the spectral bands and the distribution of the LU/LC class patterns. The data acquired in the infrared band showed better improvement over the visible band.

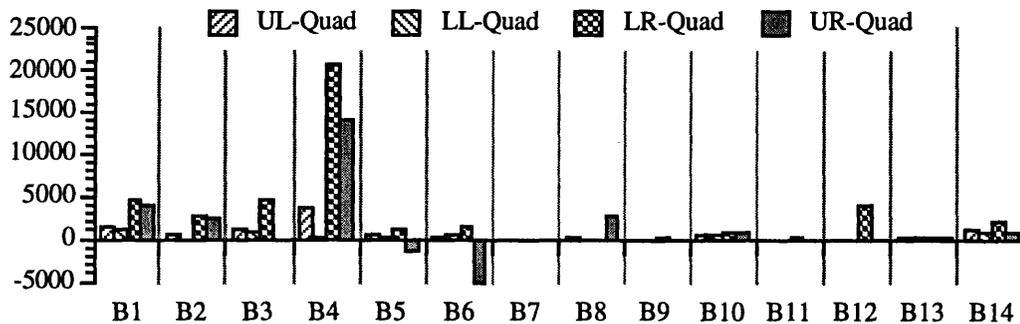


Figure 3.—Difference of sum of squares of the residuals between the OLS and the autocorrelated error model.

### CONCLUSION

The AVHRR data have autocorrelated errors. Moran's I statistics and correlation coefficients among neighboring pixels in different spatial pattern clearly demonstrated the spatial association of the AVHRR data. The spatial autocorrelation parameters were evaluated, and were corrected separately for each spectral band. The magnitude and the sign of the spatial autocorrelation parameters were dependent on many factors: size of the pixel, the spectral band of the image,

direction of the autocorrelation pattern and the shape and size of the region. Wilson (1992) observed lower autocorrelation with the adjacent diagonal pixels than with the adjacent vertical and horizontal pixels. In this study also, models using diagonal pixels for the autocorrelation patterns showed the least autocorrelation. The Rook and Bishop models further confirmed this observation. The decrease of the sum of squares of residuals in the autocorrelated error model demonstrated the advantage in accounting for the autocorrelated errors. The magnitude of the improvement certainly depends on the spectral band, acquisition period, and the distribution of the LU/LC class patterns. This study identified the advantage of analyzing the effects of autocorrelated errors in unmixing mixed pixels using simulated data with adequate background knowledge. The background knowledge is important to understanding and interpreting the effects of autocorrelated errors in unmixing mixed pixels.

## REFERENCES

- Belward, A. S. (1992). Spatial Attributes of AVHRR Imagery for Environmental Monitoring. International Journal of Remote Sensing, 13(2):193-208.
- Campbell, N.A., and H.T. Kiiveri, (1988). Neighbor Relations and Remotely Sensed Data, Report No. Internal Report: Wembley 6014, Western Australia: Division of Mathematics and Statistics, CSIRO, pp. 23.
- Ediriwickrema, D.J. (1995). Modeling and Analysis of AVHRR Data for Biogenic Emission Inventory System (BEIS). Ph. D. Dissertation. Raleigh: North Carolina State University, Raleigh, NC 27695.
- Haining, R. (1990). Spatial Data Analysis in the Social and Environmental Sciences. First ed. New York: Cambridge University Press.
- Isaaks, E.H., and R.M. Srivastava. (1989). An Introduction to Applied Geostatistics. New York: Oxford University Press.
- Iverson, L.R., E.A. Cook, and R.L. Graham. (1989). A Technique for Extrapolating and Validating Forest Cover Across Large Regions. International Journal of Remote Sensing, 10(11):1805-1812.
- Khorrarn, S., L. Allen, J. Aanstoos, J. Brockhaus, A. Sampson, H.M. Cheshire, and D.J. Ediriwickrema, (1994). Vegetation Canopy and Land Use Characterization, Report No. #68D20009: Final Report Submitted to U. S. Environmental Protection Agency, Research Triangle Park, NC 27711. Raleigh, NC 27695-7106: Computer Graphics Center, North Carolina State University, pp. 54.
- Pech, R.P., A.W. Davis, R.R. Lamacraft, and R.D. Graetz. (1986). Calibration of LANDSAT Data for Sparsely Vegetated Semi-Arid Rangelands. International Journal of Remote Sensing, 7(12):1729-1750.
- Upton, G.J.G., and B. Fingleton. (1985). Spatial Data Analysis by Example. Vol. 1. Chichester, New York: John Wiley & Sons, 2 vols.
- Wilson, J.D. (1992). A Comparison of Procedures for Classifying Remotely-Sensed Data Using Simulated Data Sets Incorporating Autocorrelations Between Spectral Responses. International Journal of Remote Sensing, 13(14):2701-2725.

# Alternative Method to Validate the Seasonal Land Cover Regions of the Conterminous United States

Zhiliang Zhu<sup>1</sup>, Donald O. Ohlen<sup>1</sup>, Raymond L. Czaplewski<sup>2</sup>,  
and Robert E. Burgan<sup>3</sup>

---

**Abstract.**— An accuracy assessment method involving double sampling and the multivariate composite estimator has been used to validate the prototype seasonal land cover characteristics database of the conterminous United States. The database consists of 159 land cover classes, classified using time series of 1990 1-km satellite data and augmented with ancillary data including terrain, climate, and ecological definitions. Reference data from more than 2,200 1 km<sup>2</sup> plots were independently collected; data included descriptions of dominant land cover types. Validity of the database is presented in terms of conditional probabilities and variability estimates. This case study demonstrates the usefulness of the method in describing the reference composition of the mapped categories, handling different classification schemes between the database and reference, and comparing various regrouping complexities. The validation results support using the database as a vegetation baseline in large scale environmental simulation models.

---

## INTRODUCTION

Data acquired by the National Oceanic and Atmospheric Administration's advanced very high resolution radiometer (AVHRR) represent one of the most important data sources for global change research (Townshend 1994). The daily coverage of a large area and 1-km spatial resolution (at nadir) of AVHRR allow users to effectively and efficiently map and study vast areas, at scales from continental to global. An example is a recent prototype land cover characteristics database for the conterminous United States (Brown and others 1993, Loveland and others 1993, 1995).

The conterminous U.S. (CONUS) land cover characteristics database, produced by the U.S. Geological Survey's EROS Data Center, in cooperation with the University of Nebraska-Lincoln, defines 159 seasonally distinct classes.

---

<sup>1</sup>Senior Scientist, Hughes STX Corporation, EROS Data Center, Sioux Falls, SD 57198.

<sup>2</sup>Mathematical Statistician, USDA Forest Service, Rocky Mountain Forest and Range Experiment Station, Fort Collins, CO 80526.

<sup>3</sup>Research Forester, USDA Forest Service, Intermountain Research Station, Missoula, MT 59807.

Vegetation seasonal properties, or phenology, are defined by a time series of the normalized difference vegetation index (NDVI), a transformation of the first two AVHRR spectral bands. Each class in the CONUS database exhibits a unique combination of seasonal characteristics and landscape pattern, such as vegetation, climate, and elevation. However, the concept and use of polythetic regionalization (Loveland and others 1995) imply that the classes consist of a mixture of land cover types within each class.

The CONUS land cover characteristics database has been widely used in large-scale environmental simulation modeling (Steyaert and others 1994). The concept and methods of the prototype database are also being used to produce similar land cover characteristics databases for the major continents of the world (Loveland and others 1995). Because of the current utility and the future applications, validation of the CONUS database is an important issue. Loveland and others (1995) discussed the importance of validation and various validation plans, including a study reported by Merchant and others (1994). Preliminary results from that study indicated that the database closely represented land cover regions of the conterminous United States. The objective of this research is to test an alternative method for validating the CONUS database, as presented below.

## ACCURACY ASSESSMENT ALTERNATIVES FOR CONUS

Accepted conventional methods for accuracy assessment, as reviewed in Congalton (1991), require that reference data and the remote sensing product in question share a common classification scheme, i.e., have the same number of corresponding classes. Positional accuracy for the map product is assessed with a square error matrix that shows percent sample points correctly classified, where the sample points are geographically coincident. The construction of error matrices also allows calculation of the kappa statistic (Congalton 1991).

The error matrix and kappa statistic approach is routinely used to validate land cover classifications with moderate to fine ground resolutions. This method has not been effective for coarse resolution products, such as those that are based on AVHRR data, for several reasons. First, the 1 km<sup>2</sup> pixel on the ground often contains a mosaic of different cover types; finding corresponding reference classes at this resolution and scale is difficult. Second, requirements of the error matrix restrict the use of reference data acquired under a classification scheme different from the one used for the satellite map. Third, image geometry is complex and identification of ground samples is difficult or speculative. Additionally, the original design philosophy (Loveland and others 1991) of the CONUS database was a flexible database of land characteristics, classes were meant to be translated or combined in different ways to meet the needs of a wide range of applications (Steyaert and others 1994). A single reference data set would be insufficient and difficult for comparing at all levels of complexities.

The consideration in this study for validation of the CONUS database was that

the reference data and validation method would be designed to describe reference land cover compositions of the mapped land cover classes, thereby allowing evaluation of the database at different levels and providing measures of map data quality to users. The method needs to be flexible and practical to permit use of a common reference for different classification schemes and to incorporate multiphase, multistage designs for improved efficiency.

The alternative method used in this study to validate the CONUS land cover characterization database is that of the multivariate composite estimator (Czaplewski 1992). The formulation for this approach is given in detail in Czaplewski (1992, 1994, 1995); a brief overview follows.

Assume that, as a special case of multiphase sampling, a set of satellite-classified (map) pixels and reference points is obtained. The proportions of each category in the sampled map and reference data are expressed with their respective state vectors and estimated covariance matrices. The multivariate composite estimator is a linear transformation matrix of the independent map and reference state vectors, with a covariance matrix calculated as the linear function of map and reference covariance matrices.

From the multivariate composite estimator, two very useful vectors are derived. The first is a vector of joint probabilities (i.e., a vectorized contingency table) showing unbiased estimates of proportions of the map and reference categories occurring together (Czaplewski 1992). The second is a nonlinear transformation vector of estimated conditional probabilities, showing estimated true ground proportions given each map category. The approximate variance estimator for the conditional probabilities is given in Czaplewski (1994).

The important features of the above work are that it provides a means of statistically describing reference composition of coarse resolution map categories, and the descriptions can be done at various levels of detail with the same reference data set. The approach accommodates the situation where different classification schemes are used by the remote sensing product and the reference data (Czaplewski 1992), and it allows for more complex sampling designs (Czaplewski 1994) beyond the simple random sample of point plots required by the conventional method (Congalton 1991).

The U.S. Forest Service has produced a software program based on this approach (Williams and Beach 1995). The program produces the joint probability vector, the conditional probability vector, and the confidence intervals calculated using the estimated variance and the beta-binomial distribution (Czaplewski 1995).

## **SAMPLING PROTOCOL AND REFERENCE PLOTS**

Congalton (1991) states that the sample size for reference data used to assess accuracies of remote sensing products is often dictated by practical considerations. In the case of the CONUS database, the study area (conterminous United States) is very large, the number of classes (159) is high, and the AVHRR spatial

resolution makes reference data expensive. Because of these constraints, the sampling design described below was a process evolving between the original design (Burgan and others 1993) and practical compromises at later stages.

A two-stage, stratified sampling design (Burgan and others 1993) was used to select 698 primary sampling units (PSU) and 3,500 secondary sampling units (SSU). The PSU's were USGS 7.5 minute quadrangle maps and the SSU's were 1-km<sup>2</sup> field plots corresponding to 1-km<sup>2</sup> pixels in CONUS database. A decision was made not to sample classes with fewer than 5,000 pixels, as well as urban areas and water. Field work was coordinated by the U.S. Forest Service (Burgan and others 1993) and carried out by personnel from several Federal and State agencies. Because of various difficulties, the actual number of SSUs measured as reference materials was 2,284. This accounted for a 0.03 percent sample intensity for 7.6 million pixels, from 129 classes, in the database. According to the procedure for reference data, each 1-km<sup>2</sup> SSU selected in a PSU was transferred to aerial photographs, interpreted, and then field-checked using a field form designed for the project. More than 100 reference variables were collected or calculated, including plot coordinates, land cover composition, and primary and secondary vegetation cover types or species.

Among these variables, the most pertinent to this validation were those of 17 dominant cover types (listed as the reference classification at the top of the figures 1-4). Within each of the 1 km<sup>2</sup> plots, percent values of land cover by the 17 cover types were tallied and summed to 100 percent. For example, the reference data for plot number 3 was 20-percent conifer with crown closure less than 30 percent, 20-percent conifer with crown closure greater than 60 percent, 50-percent mixed forest with crown closure less than 30 percent, and 10-percent shrub. The CONUS map classification for plot 3 was western U.S. conifer forests of evergreen needleleaf species. The 17 variables and the 2,284 reference points were used for comparisons to the CONUS database classes.

## RESULTS AND DISCUSSION

The reference data set was used to assess several categorical simplifications (groupings) of the CONUS database to validate the map for applications. Conditional probability vectors, standard errors, and confidence intervals at the 80 percent confidence level were obtained for the following simplifications: 6 land cover classes (figure 1), 12 land cover classes (figure 2), a modified Anderson level II classification (Anderson and others 1976) (figure 3), and a Biosphere-Atmosphere Transfer Scheme (BATS) (Dickinson and others 1986) translation (figure 4). Actual data for the four simplifications as well as the full 129 CONUS classes are also available from the authors.

In the figures, for each class category, conditional probabilities of the 17 reference dominant cover types are shown with 80 percent confidence intervals. For example, in figure 1, if the CONUS class is deciduous forests, the actual

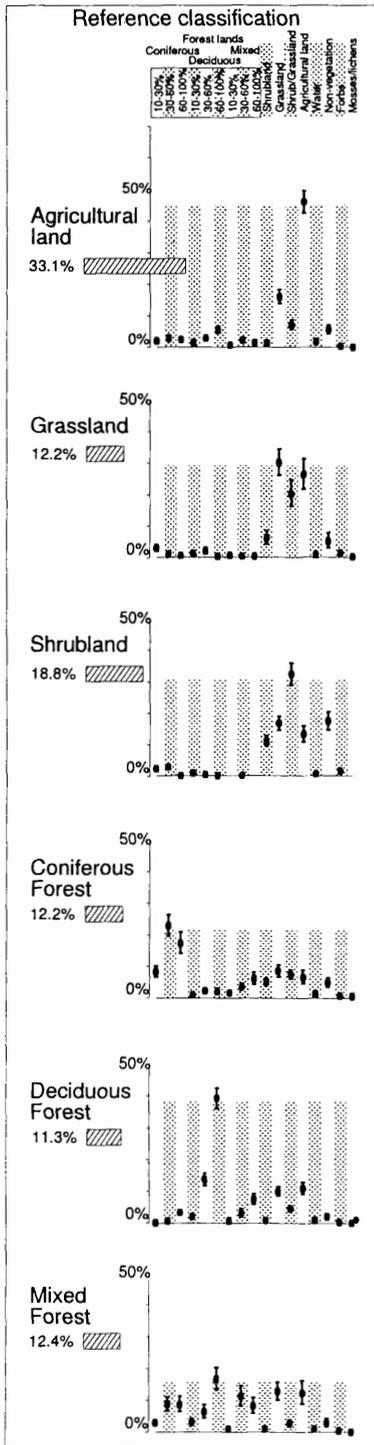


Figure 1. -- Probability of reference classification given collapsed CONUS classification system with 6 categories. Horizontal bars show prevalence of CONUS mapped categories.

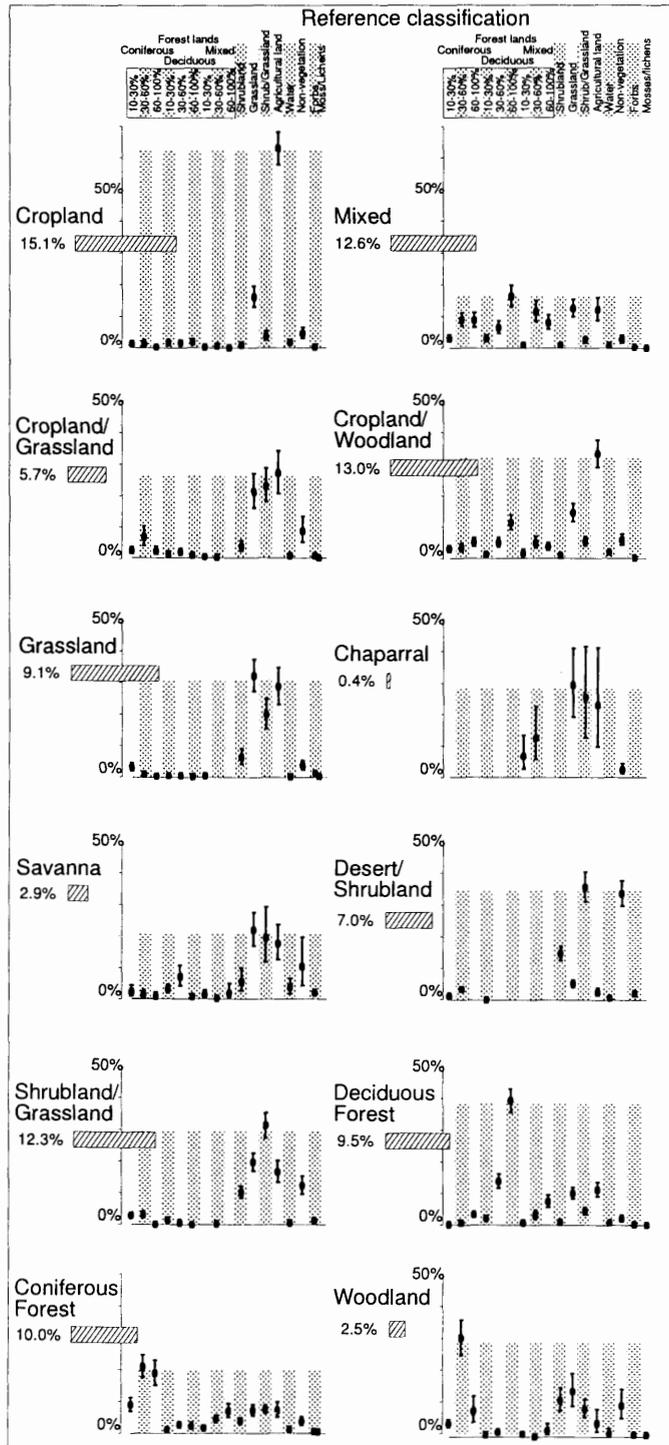


Figure 2. -- Probability of reference classification given collapsed CONUS classification system with 12 categories. Vertical lines display 80% confidence interval for the sampling error from the 698 sample quadrangles. Horizontal bars show prevalence of CONUS mapped categories.

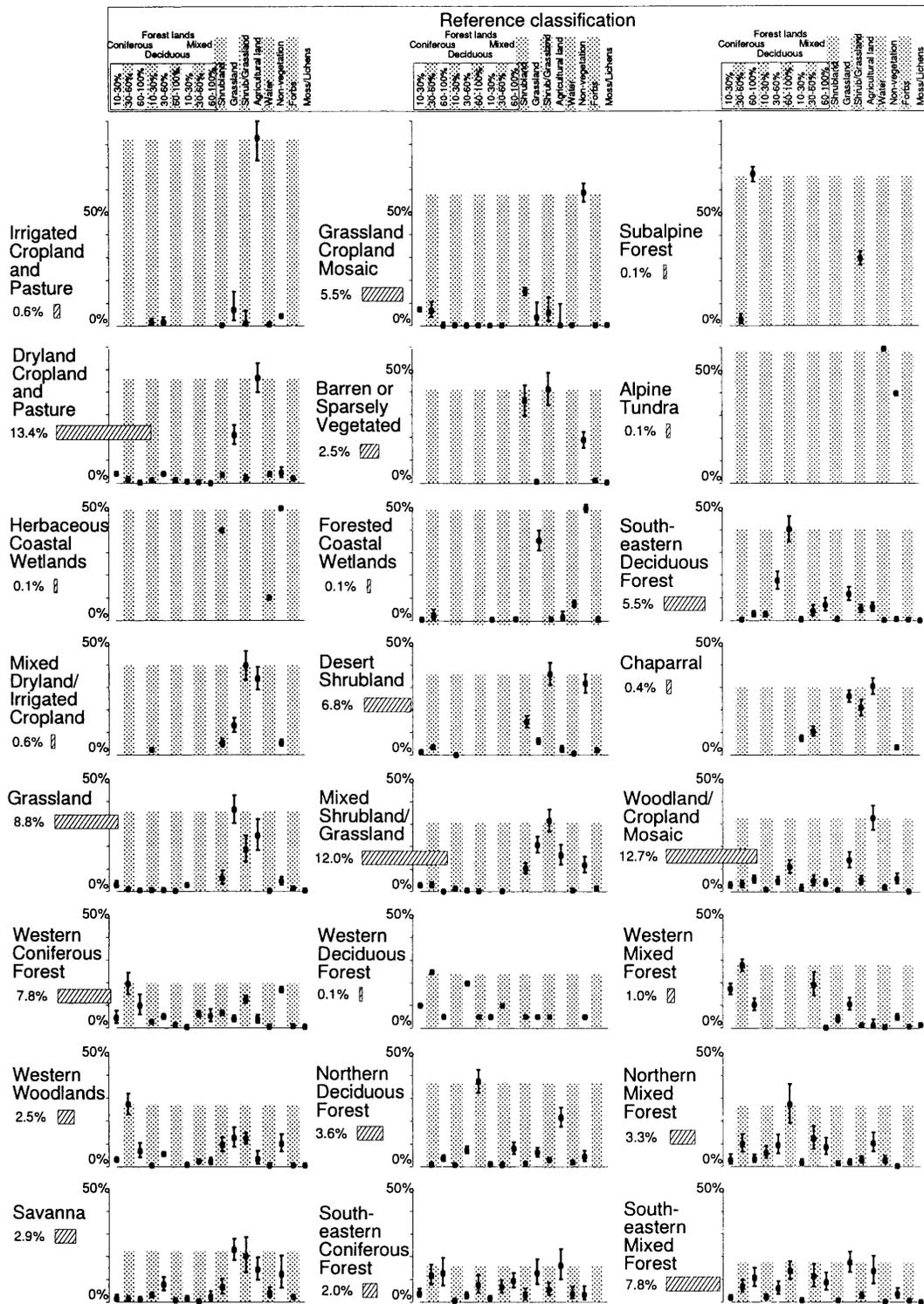


Figure 3. -- Probability of reference classification given collapsed CONUS classification system that approximates a modified Anderson Level II system. Vertical lines display 80% confidence interval for the sampling error from the 698 sample quadrangles. Horizontal bars show prevalence of mapped Anderson categories.

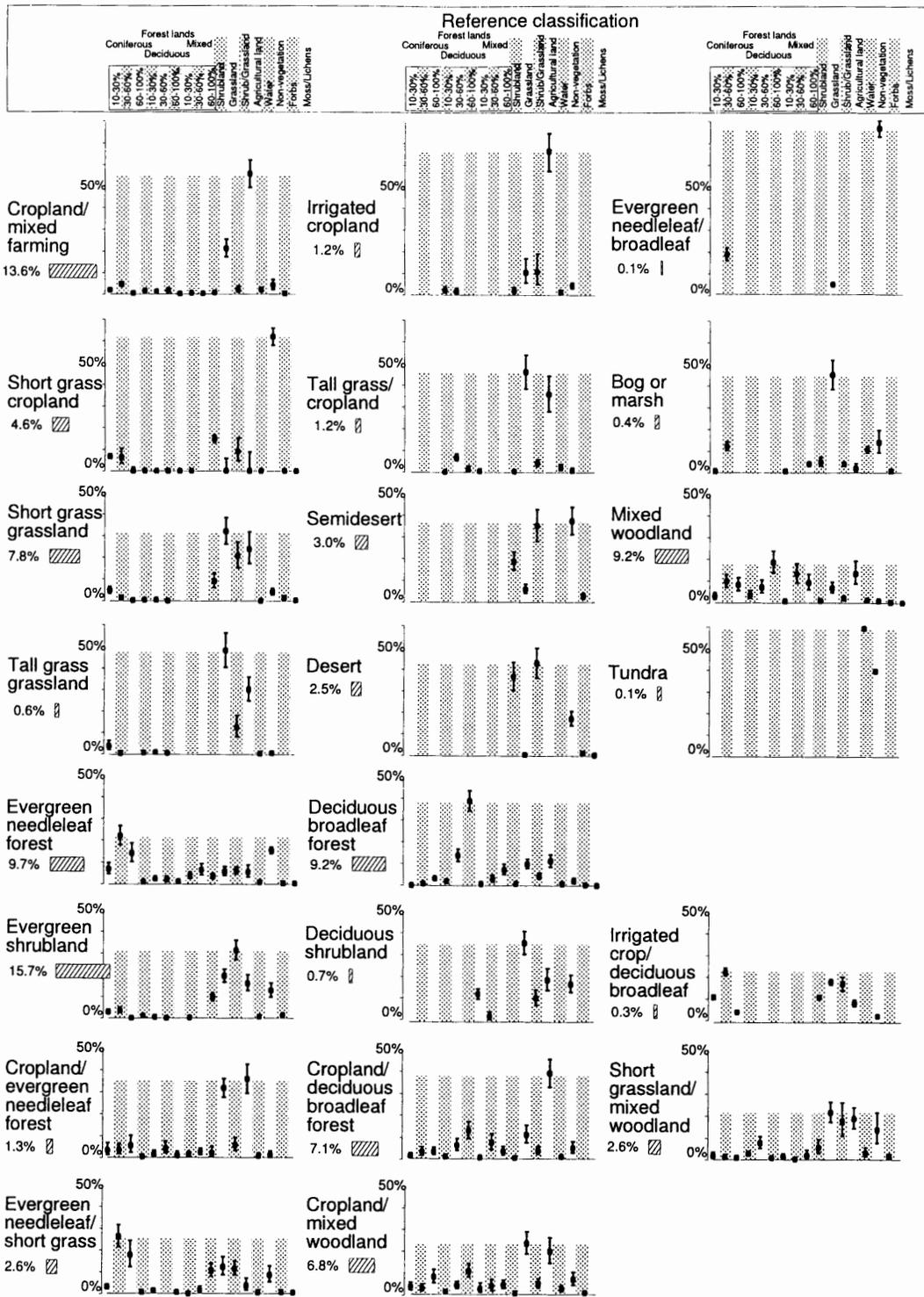


Figure 4. -- Probability of reference classification given collapsed CONUS classification system that approximates the Biosphere-Atmosphere Transfer Scheme (BATS). Vertical lines display 80% confidence interval for the sampling error from the 698 sample quadrangles. Horizontal bars show prevalence of mapped BATS categories.

cover types, according to reference field plots, are approximately 40-percent deciduous forests with crown closure greater than 60 percent, 14-percent deciduous forests with crown closure between 30 and 60 percent, 11-percent agricultural lands, 10-percent grassland, and the remaining 25-percent other cover types. In other words, the collapsed CONUS class, despite its oversimplified label as deciduous forest, is a region of pixels with distinct seasonal properties but not necessarily the conventional, mutually exclusive land cover properties (Loveland and others 1995). This seasonal region is expressed as a function of the 17-element probability vector; values of the 17 elements indicate the composition of the region. In the example above, the "deciduous" seasonal region is composed of a majority of deciduous forests (55 percent), plus agricultural and grass land cover.

The grassland class is similarly expressed in figure 1 as a seasonal region consisting of not only "pure" grass land cover (30 percent), but also a mixture of grass and shrub, crop, and forests, as well as nonvegetated land. Interpreting the results requires an understanding of the interspersion of crop and forest land cover in some areas that may resemble grassland in terms of their spectral and temporal characteristics. Misclassification of nonvegetated land caused by factors such as AVHRR data quality (low sun angle, viewing geometry, atmospheric effects) may also partially contribute to the results.

The standard error and the confidence interval are measures of sampling error for individual estimated probabilities. These variability terms can be used in large-scale environmental simulation models for calculating error propagation. It should be noted that when sample size is small for an element of the probability vector, the variance estimators may not be reliable. In this case, the variance measures should be noted but not used.

Examination of figures 1 through 4 shows the flexibility of the alternative validation method in using common reference set at different levels of CONUS simplification. Note that when CONUS classes are recoded or translated, the constituent classes are changed. For example, grassland category in the 6 class scenario (figure 1) includes savanna class, but grassland is separate from savanna in the 12 class scenario (figure 2), and it is split into short and tall grass subcategories in a BATS translation (figure 4). Compositions of deciduous forests and other categories are similarly changed. However, as shown in figures 1 through 4, their validity in terms of probabilities of ground reference and standard errors generally holds throughout the simplifications of classification detail. Deciduous classes maintain similar constituencies except the northern deciduous forest class within the Anderson translation, where the seasonal region contains a larger share of agricultural lands.

## SUMMARY

Land cover classifications and databases using coarse resolution AVHRR data

represent important inputs into environmental simulation models for global change research. Validation of AVHRR based products is critical but is often not effectively fulfilled because of the mismatch between conventional methods for accuracy assessment and characteristics of AVHRR data.

An alternative, and complementary, approach involves using conditional probability and variability terms to describe the nature and validity of the CONUS land cover characterization database. It is important to realize that all land cover classifications contain errors. As stated before, CONUS classes are different from conventional, monothetic spectral classifications in that the CONUS classes are generally dominated by one land cover type, but at the same time are comprised of other types of land cover that share similar seasonal and spectral characteristics. Conditional probabilities and error terms from this study show that the CONUS classes accurately described such landscape patterns and are generally consistent across various levels of classifications and translations.

Additional research on the validation of large area land cover mapping should look at the correlations of spatial components. Evaluating spatial errors found between the probabilities of dominant reference cover types and the various CONUS land cover mappings should provide an understanding into any regionalization of land cover seasonal characteristics.

The validation method used in this study is flexible enough to handle different classification schemes for map and reference categories and complex sampling designs, and the same reference data set can be used to evaluate map categories translated at various levels of complexity. Therefore this approach of validation and accuracy improvement for area estimates of different land cover categories is useful in land cover mapping projects of the global scale.

## ACKNOWLEDGMENTS

The research described in this article has been supported by the U.S. Environmental Protection Agency (EPA) through Interagency Agreement IAG DW14936073 to the U.S. Geological Survey. However, it has not been subjected to EPA review and therefore does not necessarily reflect the views of the agency. No official endorsement should be inferred. The Hughes STX Corporation performed this work under U.S. Geological Survey contract 1434-92-C-40004.

## REFERENCES

- Anderson, J.R., Hardy, E.E., Roach, J.T., and Witmer, R.E. 1976. A land use and land cover classification system for use with remote sensor data. *U.S. Geological Survey Prof. Paper* 964, 28 p.
- Brown, J.F., Loveland, T.R., Merchant, J.W., Reed, B.C., and Ohlen, D.O. 1993. Using multisource data in global land characterization: concepts, requirements, and methods. *Photogrammetric Engineering and Remote Sensing*, 59:977-987.

- Burgan, R.E., Hardy, C., Ohlen, D.O., Loveland, T.R., Brown, J.F., and Reed, B.C. 1993. Land characterization map evaluation, study plan no. 4401-52, U.S. Forest Service, 24 p.
- Congalton, R.G. 1991. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, 37: 35-46.
- Czaplewski, R.L. 1992. Accuracy assessment of remotely sensed classifications with multi-phase sampling and the multivariate composite estimator. *Proceedings of 16th International Biometrics Conference*, Hamilton, New Zealand, December 7-11, 1992. 2:22.
- Czaplewski, R.L. 1994. Variance approximations for assessments of classification accuracy. *Research Paper RM-316*. U.S. Department of Agriculture, Forest Service, Rocky Mountain Forest and Range Experiment Station, Fort Collins, CO. 29 p.
- Czaplewski, R.L. 1995. Assessment of classification accuracy and extent estimates for a land cover map with double sampling. Submitted to *Forest Science*.
- Dickinson, R.E., Henderson-Sellers, A., Kennedy, P.J., and Wilson, M.F. 1986. Biosphere-atmosphere transfer scheme (BATS) for the NCAR community climate model. *NCAR Technical Note NCAR/TN-275+STR*, Boulder, CO.
- Loveland, T.R., Merchant, J.W., Ohlen, D.O., and Brown, J.F. 1991. Development of a land cover characteristics database for the conterminous U.S. *Photogrammetric Engineering and Remote Sensing*, 57:1453-1463.
- Loveland, T.R., Ohlen, D.O., Brown, J.F., Reed, B.C., Merchant, J.W., and Steyaert, L.T. 1993. Prototype 1990 conterminous United States land cover characteristics data set CD-ROM, USGS CD-ROM Set 9307.
- Loveland, T.R., Merchant, J.W., Reed, B.C., Brown, J.F., and Ohlen, D.O. 1995. Seasonal land cover regions of the United States. *Annals of the Association of American Geographers*, 85(2):339-355.
- Merchant, J.W., Yang, L., and Yang, W. 1994. Validation of continental scale land cover data bases developed from AVHRR data. *Proceedings of the Pecora 12 Symposium on Land Information from Space-Based Systems*, American Society of Photogrammetry and Remote Sensing, Bethesda, MD, pp. 63-72.
- Steyaert, L.T., Loveland, T.R., Brown, J.F., and Reed, B.C. 1994. Integration of environmental simulation models with satellite remote sensing and geographic information systems technologies: case studies. *Proceedings of Pecora 12 Symposium on Land Information from Space-Based Systems*, American Society of Photogrammetry and Remote Sensing, Bethesda, MD, pp 407-417.
- Townshend, J.R.G. 1994. Global data sets for land applications from the Advanced Very High Resolution Radiometer: an introduction. *International Journal of Remote Sensing*, 15(17):3319-3332.
- Williams, M.T., and Beach, D.J.C. 1995. ACAS 0.4: accuracy assessment system program manual. U.S. Department of Agriculture, Forest Service, Rocky Mountain Forest and Range Experiment Station, Fort Collins, CO. 33 p. + source code.

# Investigation of Possible Contributions NDVI's Have to Misclassification in AVHRR Data Analysis<sup>1</sup>

David L. Evans<sup>2</sup> and Raymond L. Czaplewski<sup>3</sup>

---

**Abstract.**---Numerous subcontinental-scale projects have placed significant emphasis on the use of Normalized Difference Vegetation Indices (NDVI's) derived from Advanced Very High Resolution Radiometer (AVHRR) satellite data for vegetation type recognition. In multi-season AVHRR data, overlap of NDVI ranges for vegetation classes may degrade overall classification performance. The preliminary work of this project is presented to illustrate potential differences in vegetative cover discrimination capabilities between original visible and near-infrared (NIR) channels of AVHRR data and NDVI's derived from these channels. Initial tests on AVHRR data have been performed for regions of known vegetative cover composition in Northern Mexico. The results are presented here to stimulate debate on the effectiveness of use of NDVI's for vegetation type discrimination.

---

## BACKGROUND

The U.S. Department of Agriculture, Forest Service, has developed an increasing interest in the use of Advanced Very High Resolution Radiometer (AVHRR) data for vegetation assessments of broad regions. The AVHRR sensor is mounted on National Oceanic and Atmospheric Administration (NOAA) polar-orbiting weather satellites.

Typical applications of AVHRR data for forest assessments include land cover characterization (Brown, et al. 1993; Loveland 1991), forest density analysis (Iverson, et al. 1989; Ripple 1994; Zhu 1994), and general forest distribution

---

<sup>1</sup> *Approved for publication as Proceedings Article No. FA-055-0296 of the Forest and Wildlife Research Center, Mississippi State University.*

<sup>2</sup> *Assistant Professor of Forestry, Mississippi State University, Starkville, MS.*

<sup>3</sup> *Mathematical Statistician, USDA Forest Service, Rocky Mountain Forest and Range Experiment Station, Fort Collins, CO.*

mapping (Zhu and Evans 1994; Evans, et al. 1992). The latter of these applications utilized both the original spectral data and Normalized Difference Vegetation Indices (NDVI's) in various interpretation schemes to characterize earth cover.

The accuracy of products derived from AVHRR data is difficult to assess at continental scales due to the comprehensive coverage of the data and the low resolution of each data element (1.1 km at nadir). Therefore, increased attention has been devoted to use of data analysis techniques that optimize chances for high classification success. One possible way to increase classification success is to identify the data combinations and transformations that can be used effectively in automated classification procedures. Exclusive use of some data transformations could deprive classification procedures of valuable information that is inherent in the original data. The concern identified in this paper is that NDVI's (data transformations) may forfeit useful information that could have been obtained from the original AVHRR spectral channels.

### DERIVATION AND USES OF NDVI's

The NDVI is a data transformation (Eq.1) used to characterize vegetation vigor or "greenness."

$$\text{NDVI} = (\text{NIR} - \text{VIS} / \text{NIR} + \text{VIS}) * 100 \quad [1]$$

In Equation [1], NIR is the reflectance values for the near-infrared channel (channel 2 in AVHRR) and VIS is the reflectance values for the visible channel (channel 1 in AVHRR) from satellite data.

A common approach used to interpret AVHRR data is to acquire imagery over one year's time and calculate NDVI's to document vegetative phenology. This is done to identify different or distinctive phenological sequences that can be attributed to vegetation types, crop development, and regional climate conditions.

NDVI data have been used as input to classification algorithms to identify vegetative cover types (Brown, et al. 1993; Loveland, et al. 1991). With time as a critical component in classification, it is assumed that vegetative cover types have different phenological time frames. Given the low sensitivity of the AVHRR instrument to spectral variance in 1.1 km resolution elements, this assumption of separability may not hold in the classification process.

A complicating factor in vegetation identification is that the same vegetation may exhibit a wide range of latitudinal variation in phenology. There are numerous examples of wide-ranging eastern deciduous trees that occur from the deep south to northern states. Eastern cottonwood (*Populus deltoides*, Bartr.) ranges from southern Louisiana to central Minnesota and red maple (*Acer rubrum*, L.) ranges from central Florida to southern Canada. Many of the Oaks (*Quercus* spp. ) and hickories (*Carya* spp. ) also exhibit wide geographic ranges. These wide

geographic ranges can easily mean that leaf fall or leaf flush could vary two to three months across the range for any given species. The timing of phenologic events for any given cover type must contribute to spectral reflectance pattern variation that can ultimately cause confusion in image classification.

Confusion between different cover types with the same NDVI's can be illustrated by a hypothetical example. Let us suppose that forest land in a region has spectral reflectance values of 20 and 40 for the visible and NIR channels in AVHRR data. Grassland in the same region may have values of 30 and 60 for the same two channels. Clearly these two cover types are distinguishable in the separate spectral channels but the NDVI's for each are identical (rounded value of 33). In this hypothetical but realistic case, the use of NDVI's inadvertently filters out useful information and causes confusion. This is an assumed single-date case but even if there were multiple dates of imagery, it is possible that some NDVI's would be the same for otherwise different cover classes and therefore, would degrade classification performance.

### **RESEARCH OBJECTIVE**

The basic objective of use of AVHRR data by the USDA Forest Service is to distinguish forest cover from other earth cover over broad continental-scale regions. This has been demonstrated by (Zhu and Evans 1994). This project utilized NDVI's and the original spectral data, but placed significant reliance on ancillary data and post-processing techniques to resolve conflicts in cover class identification. The concern for future uses of AVHRR NDVI's is the ability to assess earth cover in regions where little ancillary information is available. Lack of ancillary information dictates that the maximum amount of information has to be extracted from original image data. This project was initiated to evaluate information content in satellite data to better define appropriate interpretation techniques that will maximize success in cover type identification.

### **INITIAL METHODOLOGY**

The first phase of this project is to compare cover class separability with NDVI's versus the visible and NIR AVHRR channels for areas of known ground cover. Early work has concentrated on use of data from an on-going project to map the forest cover of Mexico and Central America (Lannom 1995). Landsat TM data of Northeastern Mexico were classified as agriculture (generally fallow fields), forest, and scrub land.

AVHRR data used in this initial investigation were near-cloud-free composites derived for North America by the maximum NDVI method described by Eidenshink (1992). These data sets were composed of information compiled over 10-day intervals. Those AVHRR data sets that did not exhibit clouds or substantial compositing boundaries (due to abrupt changes in reflectance values

between image dates) were chosen for initial study. This yielded 9 data sets from those available over one year that were deemed useful for the study.

The TM data were co-registered to the AVHRR data. A linked display system implemented in PCI's Imageworks<sup>4</sup> was used to locate 3 areas in each TM cover class that covered 25 AVHRR cells (75 total cells). EASI/PACE software was then used to export the AVHRR visible, NIR, and NDVI data values to a text file for analysis with the Splus statistics software package. The discriminant analysis function in Splus was used to generate discriminant correlations (measure of the ability to identify the three cover classes) for the visible, NIR, visible and NIR combined, and NDVI channels of the AVHRR data set.

## RESULTS AND DISCUSSION

The results of the discriminant analysis work are presented in Table 1. The statistical significance of these findings has not been determined because these tests represent a limited area and should be greatly expanded to cover a wider range of conditions that are assumed to occur in large AVHRR data sets.

**Table 1.—Discriminant correlations and averages for three cover classes as represented by visible NIR and NDVI channels from selected AVHRR data sets.**

AVHRR Composite	Channels			
	VIS	NIR	VIS & NIR	NDVI
April1	0.9148	0.5656	0.9839	0.9586
April2	0.6793	0.4311	0.9878	0.8900
April3	0.9458	0.6711	0.9702	0.9323
June2	0.9062	0.2806	0.9688	0.9617
July3	0.9305	0.2411	0.9459	0.8468
Sept1	0.9549	0.2290	0.9734	0.9445
Sept2	0.9297	0.2883	0.9773	0.9187
Oct2	0.8337	0.1872	0.9470	0.9175
Nov1	0.8296	0.1802	0.9486	0.8721
Average	0.8805	0.3416	0.9670	0.9158

It was interesting to note that on average, the visible channel seemed to have better discriminating capabilities for the three subject cover classes than did the NIR. One would expect that the NIR channel, which is sensitive to vegetation response, would provide better discriminating capabilities. This area of Mexico (particularly west of the mountains) is dominated by fairly arid lands. The agriculture and scrub land categories undoubtedly had significant amounts of exposed soil that influenced the visible reflectance channel values. The visible

<sup>4</sup> Mention of company or product names is for information only and does not constitute official endorsement by Mississippi State University or the USDA Forest Service.

channel generally exhibits greater contrast and variance in spectral values between vegetated and non-vegetated cover. Thus, the greater contrast in visible channel pixels significantly influenced differentiation of forest from the other categories that had more bare ground exposed to the sensor.

The NDVI channel outperformed the visible and NIR channels when used singly. It did not seem to discriminate the subject classes as well as the combined visible and NIR channels. This finding seems to support the original premise that NDVI may mask useful information found in the original channels that could be used to discriminate different vegetative cover. Additional and more rigorous testing is recommended to substantiate these findings.

### **FUTURE WORK**

Additional testing using these or modified techniques will be much broader based than this exploratory study. TM and AVHRR data sets are readily available for different locations in Central America, Mexico, and the U.S. Testing over this broad range of conditions will provide a better determination of the utility of different AVHRR channels in vegetation discrimination. The findings from this work will likely have broad application in the improvement of on-going and future studies that utilize AVHRR for global vegetation distribution assessments.

### **LITERATURE CITED**

- Brown, Jesslyn F., Thomas R. Loveland, James W. Merchant, Bradley C. Reed, and Donald O. Ohlen. 1993. Using multisource data in global land-cover characterization: concepts, requirements, and methods. *Photogramm. Engr. and Remote Sensing*. 59(6):977-987.
- Eidenshink, Jeffery C. 1992. The 1990 conterminous U.S. AVHRR data set. *Photogramm. Engr. and Remote Sensing*. 58(6):809-813.
- Evans, David L., Zhiliang Zhu, Susan Eggen-McIntosh, Pedro Garcia Mayoral, and Jose Luis Ornelas de Anda. 1992. Mapping Mexico's forest lands with advanced very high resolution radiometer. USDA Forest Service, Southern Forest Experiment Station, Research Note SO-367, 4 p.
- Iverson, L.R., E.A. Cook, and R. L. Graham. 1989. A technique for extrapolating and validating forest cover across large regions: calibrating AVHRR data with TM data. *International J. Remote Sensing*. 10(11):1805-1812.
- Lannom, Keith B. 1995. Mapping forest distributions of Mexico and Central America. *Geomatics Info Magazine*. 9(5):33-37.

- Loveland, Thomas R., James W. Merchant, Donald O. Ohlen, and Jesslyn F. Brown. 1991. Development of a land-cover characteristics database for the conterminous U.S. *Photogramm. Engr. and Remote Sensing*. 57(11):1453-1463.
- Ripple, William J. 1994. Determining coniferous forest cover and forest fragmentation with NOAA-9 very high resolution radiometer data. *Photogramm. Engr. and Remote Sensing*. 60(5):533-540.
- Zhu, Zhiliang. 1994. Forest density mapping in the lower 48 States: a regression procedure. USDA Forest Service, Southern Forest Experiment Station, Research Paper SO-280. 11 p.
- Zhu, Zhiliang, and David L. Evans. 1994. U.S. forest types and predicted percent forest cover from AVHRR data. *Photogramm. Engr. and Remote Sensing*. 60(5):525-531.

### **BIOGRAPHICAL SKETCHES**

David L. Evans is an Assistant Professor of Forestry at Mississippi State University, Starkville, MS. He completed the Ph.D. in Forest Management at Louisiana State University in 1986. His teaching and research interests are in the use of remote sensing and geographic information systems technologies for natural resource assessments. Dr. Evans was formerly with the USDA Forest Service as a Supervisory Research Forester and Team Leader for remote sensing research in the Southern Research Station.

Ray Czaplewski has spent the past 13 years working on statistical design and analysis of regional monitoring systems for forests and other environmental features, primarily the Forest Service's nationwide Forest Inventory and Analysis Program. Other accomplishments include statistical design of the 1990 and 2000 global surveys of forests by the Food and Agricultural Organization of the United Nations, and design to monitor the effectiveness of the President's 1994 Forest Plan for the States of Oregon, Washington, and California. He has specialized in linking sample surveys to remotely sensed data and geographic information systems.

# Measures of Association and Agreement for Describing Land Cover Characterization Classes

Eugene A. Fosnight<sup>1</sup> and Gary W. Fowler<sup>2</sup>

---

**Abstract.**-- Several statistical measures of association or agreement are available for comparing land cover characterizations to reference data. This study compares five statistical measures under three sampling schemes for five different class aggregations. Results show that each statistical measure contributes to the overall view of the information. Aggregated land cover characterization classes are less successful in predicting the reference classes. Sample design and size can cause biased estimators.

---

## INTRODUCTION

Confidence in land cover characterizations derived from remotely sensed images is acquired through a thorough description of the characterizations. This paper seeks to describe measures of agreement and association suitable for quantifying the relationship between ground-sampled or photointerpreted data and characterizations of remotely sensed images. This objective consists of three tasks: (1) describe a subset of measures, (2) analyze the distribution of each measure under one simple and two stratified random sampling schemes, and (3) discuss issues of interpretability associated with each statistic.

For this study the U.S. Land Cover Characterization database was compared to a photointerpreted land use and land cover data set. The measures of association and agreement calculated were Cohen's Kappa, Cramer's V, Guttman's Proportional Reduction in Error (PRE), Goodman and Kruskal's Proportion of Explained Variance (PEV), and Percent Correctly Classified (PCC).

## MEASURES OF ASSOCIATION AND AGREEMENT

Many measures of association and agreement are available for quantifying the cross-classified tables that are created by cross-tabulating classified images and reference data. Within the remote sensing literature, these tables are often referred to as "confusion tables" or "error matrices" (Czaplewski and Catts 1992, Story and Congalton 1986).

A few general concepts should be defined. The probability for row  $i$  and column  $j$  is  $p_{ij}$ . The estimated value of  $p_{ij}$  is  $\hat{p}_{ij} = x_{ij}/N$  where  $x_{ij}$  is the cell

---

<sup>1</sup>Cartographer, Hughes STX, UNEP/GRID, EROS Data Center, Sioux Falls, SD. Work performed under U.S. Geological Survey contract 1434-92-C-40004.

<sup>2</sup> Biometrician, School of Natural Resources and Environment, University of Michigan, Ann Arbor, MI.

count. The expected value,  $m_{ij}$ , of  $x_{ij}$  given no information other than the sample marginal frequencies,  $x_{i+}$  and  $x_{+j}$  is  $m_{ij} = p_{i+}p_{+j}$ , where  $p_{i+}$  are the row marginal probabilities and  $p_{+j}$  are the column marginal probabilities.

### Measures based on Chi-Square

Several measures of association are derivatives of Pearson's Coefficient  $\Phi^2$  of Mean-Square Contingency (Eq 1). Cramer suggested modifications to restrict the range to between 0 and 1 (Bishop et al. 1975, Conover 1980). The statistical measures are

$$\Phi^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(p_{ij} - p_{i+}p_{+j})^2}{p_{i+}p_{+j}} \text{ and Cramer's } V = \left\{ \frac{\Phi^2}{\min[(I-1), (J-1)]} \right\}^{1/2}, \quad [1]$$

where the maximum achievable value of  $\Phi^2$  is  $\min[(I-1), (J-1)]$ .

As  $V$  depart from zero, the measures are interpreted as becoming less independent.

### Proportional Reduction in Error

Goodman and Kruskal (1979) discussed the PRE ( $\lambda$ ) measure based on a probabilistic model (Eq. 2). Two probabilities are compared in calculating PRE. First, the probability of an error in correctly guessing the column by chance is determined by calculating the most frequently occurring column. This probability is  $1 - p_{+m}$ , where  $p_{+m} = \max(p_{+1}, p_{+2}, \dots, p_{+J})$ .

Second, the probability of correctly guessing the column given rows is determined by calculating which rows occur most frequently with each column. This probability is

$$1 - \sum_1^I p_{im}, \text{ where } p_{im} = \max(p_{i1}, p_{i2}, \dots, p_{iJ}).$$

$$\lambda_{CIR} = \frac{(1 - p_{+m}) - (1 - \sum_1^I p_{im})}{1 - p_{+m}} \text{ and } \hat{\lambda}_{CIR} = \frac{\sum_{i=1}^I x_{im} - x_{+m}}{N - x_{+m}} \quad [2]$$

If the variables are statistically independent, PRE will be zero; however, if PRE is zero, the variables are not necessarily statistically independent. The predictive interpretation of PRE distinguishes it from Chi-square-based measures.

### Proportion of Explained Variance

After Bishop et al. (1975), Guttman's PEV ( $\tau$ ) can be explained as the difference between predicting the row given only the marginal probabilities of the columns and predicting the column given the conditional proportions of the rows (Eq. 3).

PEV can be cast into analysis of variance terminology, where the total variation is defined as:

$$TSS = \frac{N}{2} - \frac{1}{2N} \sum_{j=1}^J x_{+j}^2$$

This definition of total variation meets the criterion of equaling zero if and only if all counts are in the same column, and it is maximized when the marginal frequencies of the columns are uniformly distributed.

The total variation is split into between (BSS) and within (WSS) class components

$$WSS = \frac{N}{2} - \frac{1}{2} \sum_{i=1}^I \frac{1}{x_{i+}} \sum_{j=1}^J x_{ij}^2, \text{ and } BSS = TSS - WSS = \frac{1}{2} \sum_{i=1}^I \frac{1}{x_{i+}} \sum_{j=1}^J x_{ij}^2 - \frac{1}{2N} \sum_{j=1}^J x_{+j}^2.$$

PEV is then set equal to the ratio of the between class component and the total variation.

$$\hat{v}_{CIR} = \frac{BSS}{TSS}. \quad [3]$$

The measure PEV for the column given the rows can be described as the percentage of the variation in the columns that can be explained knowing the rows.

### Measures of Agreement

Measures of agreement have more restrictive class specifications than do measures of association, but they also provide more powerful tests if the class specifications can be met. The cross-classified table must have identical classes in the same order for both rows and columns. This restriction gives special meaning to the diagonal of the table.

PCC, the simplest measure of agreement, is calculated by dividing the sum of the diagonal cells (all samples correctly classified) by the total number of samples (Eq. 4).

$$PCC = \sum_{i=1}^I p_{ii} \text{ and } P\hat{C}C = \sum_{i=1}^I x_{ii} / N. \quad [4]$$

This is the most common statistic for assessing accuracy in the remote sensing literature. PCC has a highly intuitive interpretation.

Cohen (1960) proposed a measure of agreement, Kappa, (Eq. 5) that adjusts PCC for the probability of chance agreement. Chance agreement is the probability of agreement given only the sample marginal proportions.

$$K = \frac{\sum_{i=1}^I p_{ii} - \sum_{i=1}^I m_{ii}}{1 - \sum_{i=1}^I m_{ii}} = \frac{\sum_{i=1}^I p_{ii} - \sum_{i=1}^I p_{i+}p_{+i}}{1 - \sum_{i=1}^I p_{i+}p_{+i}} \text{ and } \hat{K} = \frac{N \sum_{i=1}^I x_{ii} - \sum_{i=1}^I x_{i+}x_{+i}}{N^2 - \sum_{i=1}^I x_{i+}x_{+i}} \quad [5]$$

Kappa equals 1 if there are no off-diagonal counts. With the exception of perfect agreement Kappa will always be lower than PCC. Unfortunately, as noted by Card (1982) and Hay (1979), PCC can also be a highly biased estimator.

### Calibration

The estimated probability of each cell in the cross-classified table (Eq. 6) can be calibrated by incorporating the known population marginal frequencies:

$$\hat{p}_{ij}^* = \frac{X_{.j} \hat{p}_{ij}}{N \hat{p}_{.j}}, \text{ where } X_{.j} = \sum_{j=1}^J X_{ij} \text{ are the population marginal frequencies.} \quad [6]$$

Calibrated estimates of the cell counts (Eq. 7) can then be calculated from the cell probabilities and the known finite population is

$$x_{ij}^* = \hat{p}_{ij}^* N. \quad [7]$$

Card (1982) and Czaplewski and Catts (1992) described in detail calibration techniques for deriving calibrated estimates for the areas of reference classes given known population frequencies of the modeled classes.

The measures of association and agreement, Cramer's V, Guttman's PRE, Goodman and Kruskal's PEV, PCC, and Cohen's Kappa form a representative sample of the measures that could contribute toward assessing the accuracy of remotely sensed classifications.

## EXPERIMENTAL DATA

The study area is the conterminous United States. The U.S. Land Cover Characterization database is compared to the reference U.S. Geological Survey (USGS) Land Cover and Land Use digital data. The U.S. Land Cover Characterization database is derived from Advanced Very High Resolution Radiometer (AVHRR) images (Loveland et al. 1991, Brown et al. 1993). The USGS Land Cover and Land Use digital data are derived from photographs at scales smaller than 1:60,000 (Anderson 1976, USGS 1990).

The two classifications were grouped into several alternative aggregations. Three levels of the Anderson classification scheme were created for the reference data set: (1) the 37-class Anderson level II (lulc), (2) a 14-class modified Anderson that retains the classes of natural vegetation, range (mixed, shrub and brush, and herbaceous), forest (deciduous, evergreen, and mixed), and wetlands (forested and nonforested) at Anderson Level II, while collapsing the remaining classes to Anderson Level I (lulcMA), and (3) an 8-class Anderson Level I (lulcI).

The level I classification was determined by Anderson to be the class resolution attainable by Landsat Multispectral Scanner scale classifications. The Level II classification was considered attainable using small-scale (less than 1:60,000) aerial photographs. The detailed temporal information available with AVHRR images (daily coverage) and the ancillary information provide the possibility of attaining at least level I class resolutions.

Three classifications were created from the land cover characterization: (1) the ungrouped classes (lcc), (2) an aggregation corresponding to the lulcI (lccI) classes, and (3) an aggregation corresponding to the lulcMA (lccMA) classes.

The classifications of the land cover characterization classes and reference land use land cover classes were cross-classified into five tables; lcclulc, lcclulcMA, lcclulcI, lccMAlulcMA, lccIulcI. The first three cross-classified tables (lcclulc, lcclulcMA, and lcclulcI) are designed to show how well the full set of land cover characterization classes can predict decreasing levels of class resolution within the Anderson classification. The square cross-classified tables use Cohen's Kappa and PCC statistics for measuring agreement between the variables.

## RESULTS

The parameters provide a baseline for comparison with simulation results. When lcc is compared to the three lulc aggregations, the ability of lcc to predict lulc, as measured by  $PRE_{lulc|lcc}$ , increases as the number of classes in lulc decreases, likewise, the proportion of explained variance, as measured by  $PEV_{lulc|lcc}$ , and the lack of independence, as measured by Cramer's V, also increase. The least amount of predictive or explanatory power exists when lcc is aggregated to the modified Anderson classification, where the class resolution of the land cover characterization classes are stretched to the limit. When the data sets are aggregated to class resolution more appropriate to the AVHRR satellite, as in  $lcc|lulcI$ , the measures approach their maximum.

The results of the Monte Carlo simulations for simple random sampling, stratified random sampling with probability proportional to size, and stratified random sampling with equal probabilities are shown in Figures 1, 2, and 3, respectively. The Monte Carlo simulations consisted of a thousand repetitions calculated for each of the measures for random samples selected without replacement from the full data set. The population parameters are shown on each of the figures.

All of the measures have inflated values for small sample sizes extracted from the large rectangular tables. Whether this is caused by the large number of cells (and large number of empty cells) or the rectangular nature of the tables is unknown; however, the effect is least pronounced in table  $lcc|lulcI$ , which is the most rectangular and has the fewest empty cells. Cramer's V shows some bias for all tables at small sample sizes.

For simple random sampling and stratified random sampling with probability proportional to size, the distributions are very similar. Stratified random sampling with probability proportional to size shows slightly larger biases for small sample sizes. The stratified sampling by lcc class enforced a more even distribution of samples across lulc classes, causing fewer classes with a marginal frequency of zero.

Under stratified random sampling with equal probabilities, PRE, PEV, and Cramer's V are consistently overestimated, while for Kappa and PCC the measures are consistently underestimated (Figure 3).

Equal probability sampling results in small classes being over-sampled and large classes being under-sampled. For large percent samples, the small classes are exhaustively sampled, and hence cannot achieve the target percentage. Certainly in the case of a 10 percent sample, all 159 lcc classes cannot be sampled at 10 percent of the total. The very smallest classes are nearly exhaustively sampled with the .01 percent sample.

Because equal probability sampling is not necessarily representative of the population, it becomes susceptible to differential accuracy among the classes. The net result is that the statistical measures are highly biased for even large sample sizes, and the direction of the bias depends on the statistic and per class accuracies.

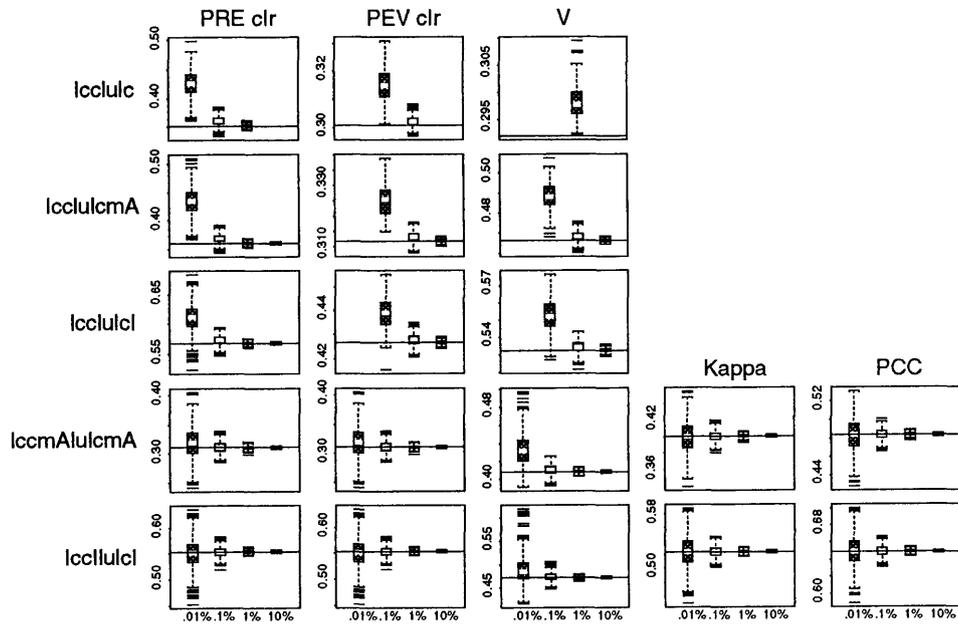


Figure 1.—Results from Monte Carlo simulation of simple random sample for measures of association and agreement. Missing boxplots result either from all zero cells for an lcc or an lulc class or from computation time too long to complete. The horizontal lines through the boxes are the population parameters.

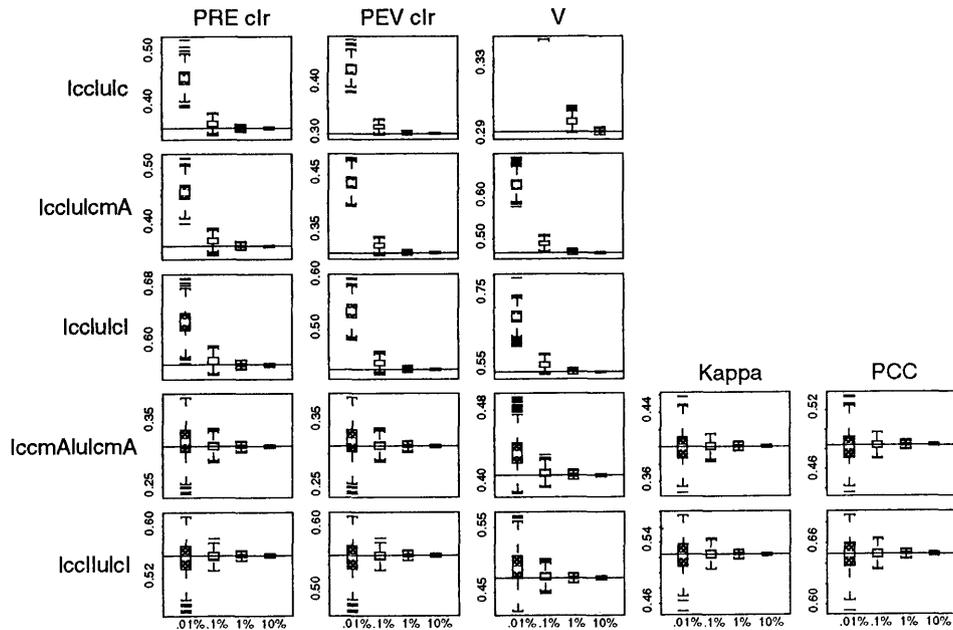
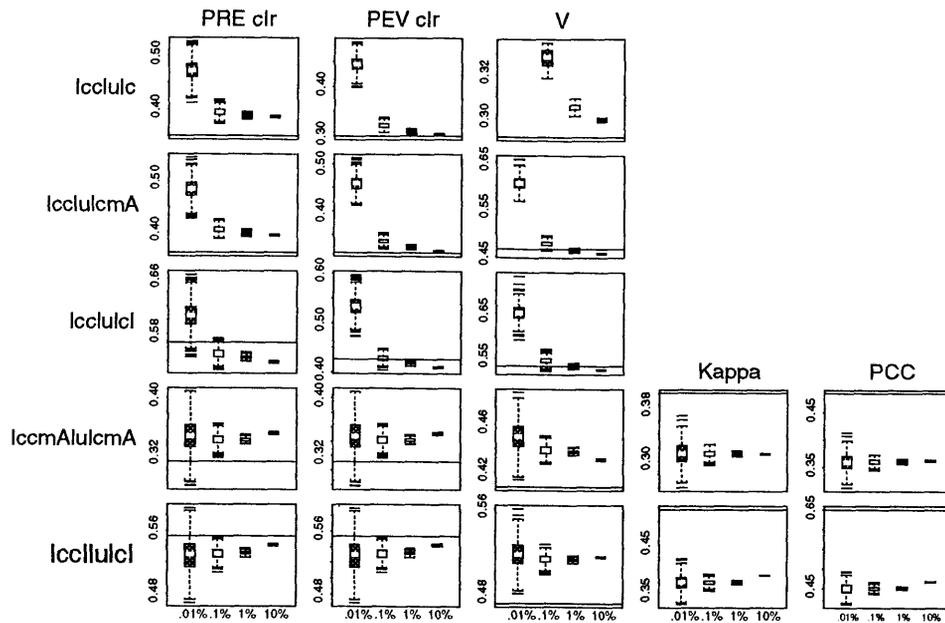


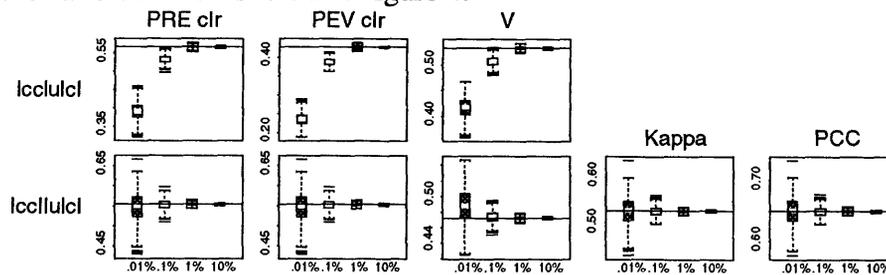
Figure 2.—Results from Monte Carlo simulation of stratified random sampling with probability proportional to size for measures of association and agreement. Missing boxplots result from all zero cells for either an lcc or an lulc class. The horizontal lines through the boxes are the population parameters.



**Figure 3.—Boxplots of results from Monte Carlo simulation of stratified random sampling with equal probability allocation for measures of association and agreement. Missing boxplots result from all zero cells for either an lcc or an lulc class. The horizontal lines through the boxes are the population parameters.**

Equal probability sampling ensures a sufficient sample size of each stratified class. This allows a sufficient description of each of the stratified, in this case lcc, classes. Under probability proportional to size, sampling some classes may be under-represented. For example, rare classes may be sampled only once, if at all, to meet strict probability proportional to size criteria.

The cell counts resulting from stratified random sampling with equal probabilities can be adjusted to better conform to cell probabilities yielded through simple random sampling. In this Monte Carlo simulation, each sample from the cross-classified tables *lcclulcl* and *lccllulcl* was selected with equal probabilities and was calibrated using the population marginal sums for the lcc classes. The results from the calibration are shown in Figure 4.



**Figure 4.—Results from Monte Carlo simulation of stratified random sampling with equal probability allocation for measures of association and agreement after calibration of cell counts. The horizontal lines through the boxes are the population parameters.**

The distributions after calibration for the cross-classified table *lccllulcl* are very similar to those from the simple random sample. For the larger rectangular table *lcclulcl*, even though the biases are removed for large sample sizes, the biases

for small samples grew to be even larger. For this type of table, the small sample sizes and large number of empty cells prevent the calibration from being well controlled. Unfortunately, these are the practical sample sizes for actual applications.

## CONCLUSION

Only three of the measures, PRE, PEV, and Cramer's V, are suitable for rectangular tables. PRE and PEV are asymmetrical predictive measures, and Cramer's V is a measure of independence. Cramer's V shows some tendency to be biased for small sample sizes, even for a small number of classes.

Both PEV and PRE were designed for ease of interpretation. Cramer's V can perhaps be used best in tandem with other measures. For example, a table can be tested for independence with Cramer's V, and then the strength of the association can be measured with PRE and PEV. For those cross-classified tables where Kappa is applicable, it provides the most robust statistic; however, PCC is the simplest to understand.

## REFERENCES

- Anderson, J.R., Hardy, E.E., Roach, J.T., and Witmer, R.E., 1976, A Land Use and Land Cover Classification System for Use with Remote Sensor Data: Washington, DC, U.S. Government Printing Office.
- Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W., 1975, Discrete Multivariate Analysis: Cambridge, Massachusetts, The MIT Press.
- Brown, J.F., Loveland, T.R., Merchant, J.W., Reed, B.C., and Ohlen, D.O., 1993, Using Multisource Data in Global Land-Cover Characterization: Concepts, Requirements, and Methods: Photogrammetric Engineering and Remote Sensing, v. 59, no. 6, p. 977-987.
- Card, D.H., 1982, Using Known Map Category Marginal Frequencies to Improve Estimates of Thematic Map Accuracy: Photogrammetric Engineering and Remote Sensing, v. 48, no. 3, p. 431-439.
- Cohen, J., 1960, A Coefficient of Agreement for Nominal Scales: Educational and Psychological Measurement, v. 20, no. 1, p. 37-46.
- Conover, W.J., 1980, Practical Nonparametric Statistics: New York, NY: John Wiley & Sons.
- Czaplewski, R.L., and Catts, G.P., 1992. Calibration of Remotely Sensed Proportion or Area Estimates for Misclassification Error: Remote Sensing of Environment, v. 39, pp. 29-43.
- Goodman, L.A., and Kruskal, W.H., 1979, Measures of Association for Cross Classifications, Springer Series in Statistics: New York, NY: Springer-Verlag.
- Hay, A.M., 1979, Sampling Designs to Test Land-Use Map Accuracy: Photogrammetric Engineering and Remote Sensing, v. 45, no. 4, p. 529-533.
- Loveland, T.R., Merchant, J.W., Ohlen, D.O., and Brown, J.F., 1991, Development of a Land-Cover Characteristics Database for the Conterminous

- U.S.: Photogrammetric Engineering and Remote Sensing, v. 57, no. 11, p. 1,453-1,463.
- Story, M., and Congalton, R.G., 1986, Accuracy Assessment: A User's Perspective: Photogrammetric Engineering and Remote Sensing, v. 52, no. 3, p. 397-399.
- U.S. Geological Survey, 1990, Land Use and Land Cover Digital Data from 1:250,000- and 1:100,000-Scale Maps: Reston, Virginia, U.S. Geological Survey.

### **ACKNOWLEDGMENTS**

The research described in this article has been supported by the U.S. Environmental Protection Agency (EPA) through Interagency Agreement IAG DW14936073 to the U.S. Geological Survey and through the National Aeronautics and Space Administration in support of the United Nations Environment Programme. However, it has not been subjected to EPA review and therefore does not necessarily reflect the views of the agency. No official endorsement should be inferred by either the U.S. Government or the United Nations Environment Programme.

### **BIOGRAPHICAL SKETCH**

Eugene A. Fosnight is a cartographer with the United Nations Environmental Programme's Global Resource Information Database (GRID), Sioux Falls, S. Dak. He graduated from Purdue University with a B.S. in 1972, from University College Swansea with a diploma in cartography in 1980 and from University of Michigan with a M.S. in remote sensing in 1992. Gene provides cartographic, remote-sensing and statistical support for GRID-Sioux Falls.

Gary W. Fowler is a Professor of biometrics with the School of Natural Resources and Environment at the University of Michigan, Ann Arbor, Mich. He graduated from the University of California with a Ph.D. in 1969. Gary's research concentrates on sequential sampling, estimation of forest stem parameters, volume-basal area ratios for use in horizontal point sampling, efficient sampling of endemic forest insect populations, development of new volume equations for commercial coniferous species, and statistical properties of species diversities.

# Forest Spatial Surveys Using the Rao-Hartley-Cochran Sampling Design

Jeffrey S. Pontius<sup>1</sup>

---

**Abstract.**--An illustration of the Rao-Hartley-Cochran (RHC) sampling strategy for a simulated forest inventory is presented. The objective of this illustration is to promote the consideration of the RHC strategy when a probability proportional to size with replacement strategy is being considered in a forest inventory. The illustration is based on a sampling frame of partitioned selection areas of the forest floor, and comparing the RHC strategy to a Hansen-Hurwitz probability proportional to size strategy.

---

## INTRODUCTION

This note presents an illustration of the Rao-Hartley-Cochran sampling strategy in the context of forest sampling when the selection units are partitioned areas of a forest floor. In forest surveys, the goal is to obtain reliable estimates of characteristics of a forest based on a sampling strategy. Naturally, one would like to use a sampling strategy that results in unbiased estimates and 'small' standard errors of those estimates. The Rao-Hartley-Cochran strategy can provide smaller standard error estimates as compared to a with replacement, probability proportional to size strategy.

### Sampling Frame

There are various approaches to sampling a forest based on particular sampling frames (de Vries 1989), which depend on the objectives of the survey. The sampling frame used in this illustration is based on Roesch, Green and Scott (1993). This sampling frame is constructed by partitioning the forest floor into mutually exclusive segments based on the (overlapping) trees' K-circles (a circle centered at the tree with a fixed radius). The resulting segments are then selected, using some randomization mechanism, with probabilities proportional to the size of the segments. The trees whose selection areas contain the randomly selected segment are included in the sample.

Following the notation in Roesch, Green and Scott (1993), let  $\{(1, \tilde{y}_1), (2, \tilde{y}_2), \dots, (N, \tilde{y}_N)\}$  denote the finite population of trees with their unit labels and

---

<sup>1</sup>Assistant Professor, Department of Statistics, Kansas State University, Manhattan, KS.

associated values of a characteristic of interest. Let  $\{\tilde{A}_1, \tilde{A}_2, \dots, \tilde{A}_N\}$  be the set of selection areas of the trees formed by their K-circles. The set of segment areas  $\{A_j, j = 1, 2, \dots, M\}$ , determined by the overlapping areas,  $\tilde{A}_i$ , or  $A_j = \tilde{A}_i$  for some nonoverlapping selection areas, forms the sampling frame.

### A REVIEW OF THE SAMPLING STRATEGIES

A brief review of the Hansen-Hurwitz and the Rao-Hartley-Cochran strategies using the above sampling frame is presented.

#### Hansen-Hurwitz Estimation

Roesch, Green and Scott (1993) constructed a sampling strategy that incorporated Hansen-Hurwitz (1943) {HH} estimation. The probability that selected segment  $j$  adds tree  $i$  to the sample is  $p_{ij} = (A_j / \tilde{A}_i) Z_{ij}$ ,

$$\text{where } Z_{ij} = \begin{cases} 1, & \text{segment } j \text{ is part of the } i^{\text{th}} \text{ tree's K-circle} \\ 0, & \text{otherwise} \end{cases},$$

and the value associated with segment  $j$  is  $y_j = \sum_{i=1}^N p_{ij} \tilde{y}_i$ . An estimator of the

parametric total  $Y = \sum_{j=1}^M y_j = \sum_{j=1}^M \tilde{y}_j$  is

$$\hat{Y}_{\text{HH}} = \frac{1}{m} \sum_{j=1}^m \frac{y_j}{p_j}, \quad (1)$$

where  $p_j = A_j / A_T$  and  $A_T$  is the area of the ground in the forest to be surveyed.

The variance of the estimator in Eq. 1 is

$$V(\hat{Y}_{\text{HH}}) = \frac{1}{m} \sum_{j=1}^M p_j \left( \frac{y_j}{p_j} - Y \right)^2 \quad (2)$$

with variance estimator

$$v(\hat{Y}_{\text{HH}}) = \frac{1}{m(m-1)} \sum_{j=1}^m \left( A_T \frac{y_j}{A_j} - \hat{Y}_{\text{HH}} \right)^2.$$

## Rao-Hartley-Cochran Estimator

Estimators based on without replacement sampling are generally more efficient (i.e., the estimators have smaller variance) than estimators based on with replacement sampling. The well-known Rao-Hartley-Cochran sampling strategy (Rao, Hartley and Cochran 1962) {RHC} applied to the segmentation sampling frame is as follows. Partition the universe  $U$  of  $M$  segments into a random grouping  $\{G_1, G_2, \dots, G_g, \dots, G_m\}$ , where  $1 \leq M_g$ , the groups are selected with equal probabilities, and  $M = \sum_{g=1}^m M_g$ . Simple random sampling without replacement

(SRSWOR) can be used to assign segments to groups by selecting a SRSWOR of  $M_1$  segments from  $M$  (for group  $G_1$ ), then selecting a SRSWOR of  $M_2$  segments from  $U - G_1$ , then selecting a SRSWOR of  $M_3$  segments from  $U - (G_1 \cup G_2)$ , ..., and, finally, assigning the remaining segments to  $G_m$ . Then, independently from each group, select one segment with probability proportional to the normed size measures  $A_j$  within each group. Specifically, those probabilities are  $p_g = p_j/P_g$ ,  $g = 1, 2, \dots, M_g$ , for segment  $j$  in group  $G_g$ , where  $p_j = A_j/A_T$  and  $P_g = \sum_{j \in G_g} p_j$ .

Let  $y_g$  and  $A_g$  be those  $m$  values of  $y_j$  and  $A_j$ , respectively, selected from group  $g$ . The unbiased RHC estimator of the parametric total  $Y$  is

$$\hat{Y}_{\text{RHC}} = A_T \sum_{g=1}^m P_g \frac{y_g}{A_g}. \quad (3)$$

Chauduri and Mitra (1992) indicated that the variance derived by Rao, Hartley and Cochran (1962) usually performs as well as an alternative variance estimator proposed by Ohlsson (1989), so the RHC variance estimator was used in the simulation illustration. Rao, Hartley and Cochran (1962) gave several group size conditions for comparing the variance of  $\hat{Y}_{\text{RHC}}$  to the variance of  $\hat{Y}_{\text{HH}}$ . In the illustration (see next section), the variance minimizing condition  $M_1 = M_2 = \dots = M_m = M/m$  applies, so the form

$$V(\hat{Y}_{\text{RHC}}) = \left(1 - \frac{m-1}{M-1}\right) V(\hat{Y}_{\text{HH}}),$$

with variance estimator

$$v(\hat{Y}_{\text{RHC}}) = \frac{1}{(m-1)} \left( \frac{M-m}{M} \right) \sum_{j=1}^m \left( A_T \frac{y_j}{A_j} - \hat{Y}_{\text{RHC}} \right)^2 \quad (4)$$

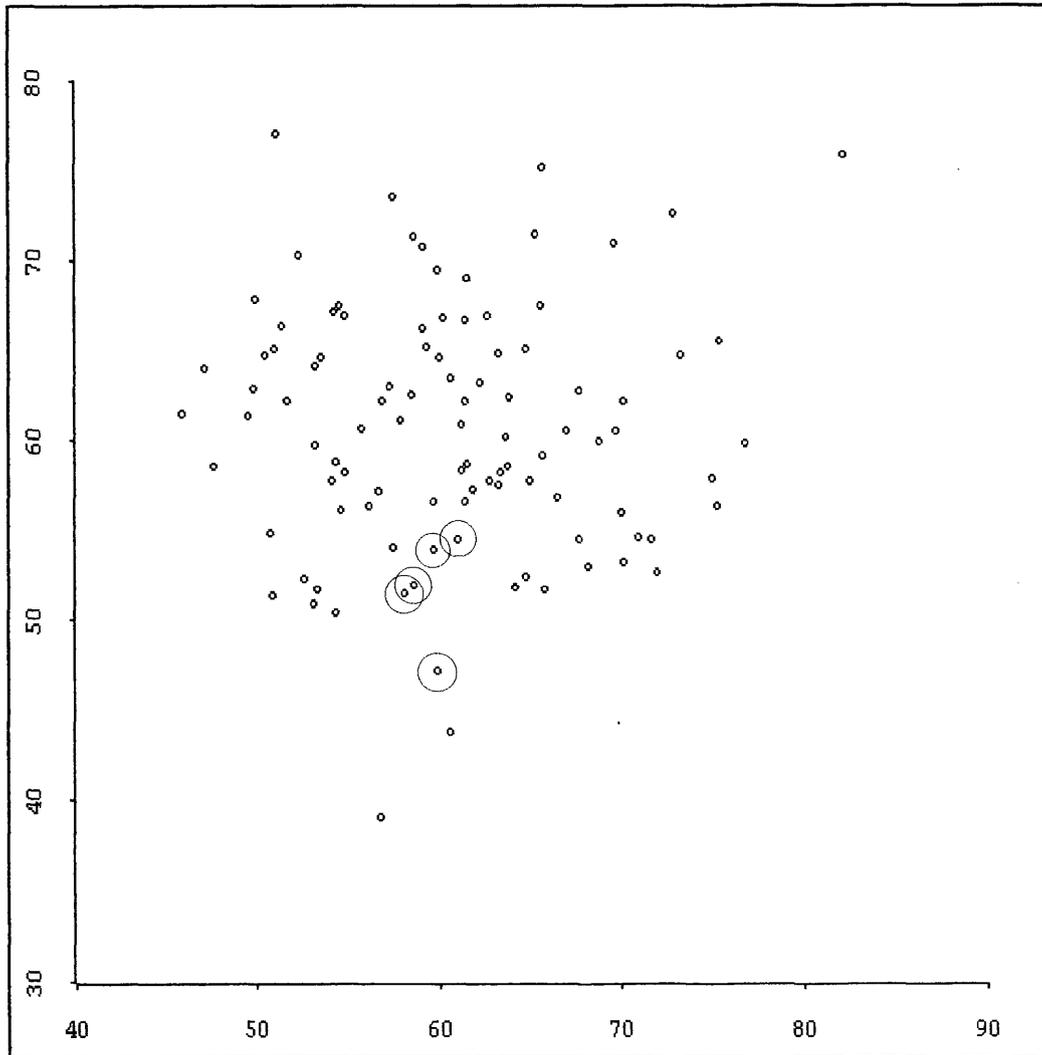
was used. Note that there has been some discussion about how well the efficiency of the RHC strategy holds as compared to some with replacement sampling strategies (Deshpande 1984, Singh and Kishmore 1975). The simulation illustration provides an additional view on this issue.

### AN ILLUSTRATIVE COMPARISON OF THE STRATEGIES

A small simulation study was designed to investigate some of the empirical properties of the RHC and the HH strategies for the above segmentation sampling frame. A dispersion pattern of  $N = 100$  "trees" was obtained by generating the coordinates  $(x, y)$  using independent random values from Poisson(60) + Uniform(0, 1) (figure 1). The volume ( $m^3$ ) of each tree was generated using independent normal(8, 2) random values based on the timber volumes in an example of de Vries (1989, p. 77). Note that the use of the Poisson and normal distributions give 'well-behaved' characteristics to the data. The segmentation sampling frame was laid out as described above, the  $M = 156$  areas  $A_j$  computed, and the pairs  $(A_j, y_j)$ , entered into a data list. Here,  $A_T$  is the sum of the  $A_j$ . Using the  $M/m$  criterion with  $M = 156$ , the following set of nontrivial pairs  $\{(m, M_g)\}$  was available for RHC sampling  $\{(2, 78), (4, 39), (6, 26), (12, 13), (26, 6), (39, 4), (78, 2)\}$ . The RHC and HH strategies were simulated 1000 times for each sample size  $m = 6, 26$  or  $78$  using XLISP-STAT (Tierney 1991). The estimators in Eq.'s 1, 2, 3, and 4 were used in computations. Boxplots of the estimates of total tree volume,  $\hat{Y}_{\text{RHC}}$  or  $\hat{Y}_{\text{HH}}$ , and estimates of their standard errors,  $\left[ v(\hat{Y}_{\text{RHC}}) \right]^{\frac{1}{2}}$  or  $\left[ v(\hat{Y}_{\text{HH}}) \right]^{\frac{1}{2}}$ , are used to summarize the simulation results (figures 2 and 3).

Overall, both strategies resulted in the median of the estimates close to the parametric total  $\tau = 793 \text{ m}^3$ , and the RHC estimates were slightly less variable than the HH estimates (figure 2). Because both  $\hat{Y}_{\text{RHC}}$  and  $\hat{Y}_{\text{HH}}$  are unbiased estimators with respect to their sampling designs, these results were anticipated. As the sample size increased, both sets of estimates became less variable. Variability of the estimates was similar for  $m = 6$ , but the RHC strategy produced estimates with proportionately less variability as compared to the HH strategy as the sample size increased.

The estimates of the standard errors became smaller as sample size increased, and the estimates from the RHC strategy became proportionately smaller than the estimates from the HH strategy as the sample size increased. The variability in the



**Figure 1.—Geometric representation of the dispersion of “trees.” Each tree was generated by two independent random values obtained from  $\text{Poisson}(60) + \text{Uniform}(0, 1)$ . Circles illustrate the K-circles of the trees.**

RHC standard error estimates decreased proportionately relative to the HH standard error estimates as sample size increased. Thus the standard error of the RHC strategy (Eq. 2) tends to give smaller standard error estimates relative to the HH estimates (Eq. 4), although the practical difference appears minimal for the generated Poisson dispersion pattern and the normally distributed tree volumes of this illustration.

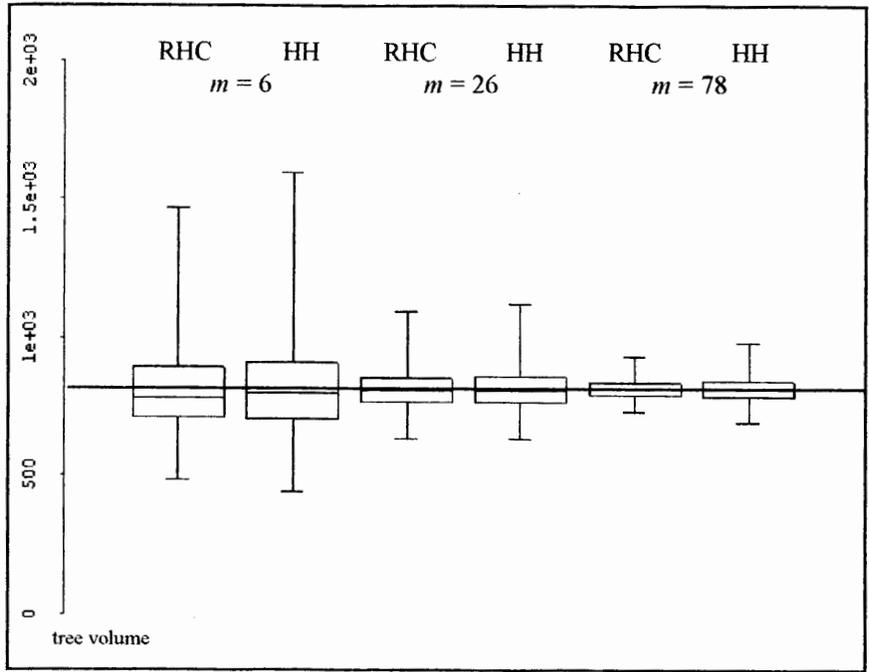


Figure 2.—Boxplots of the Rao-Hartley-Cochran (RHC) and Hansen-Hurwitz (HH) estimates of total tree volume. The total tree volume is 793.6 cubic meters, indicated by the horizontal line. Sample size is  $m$ .

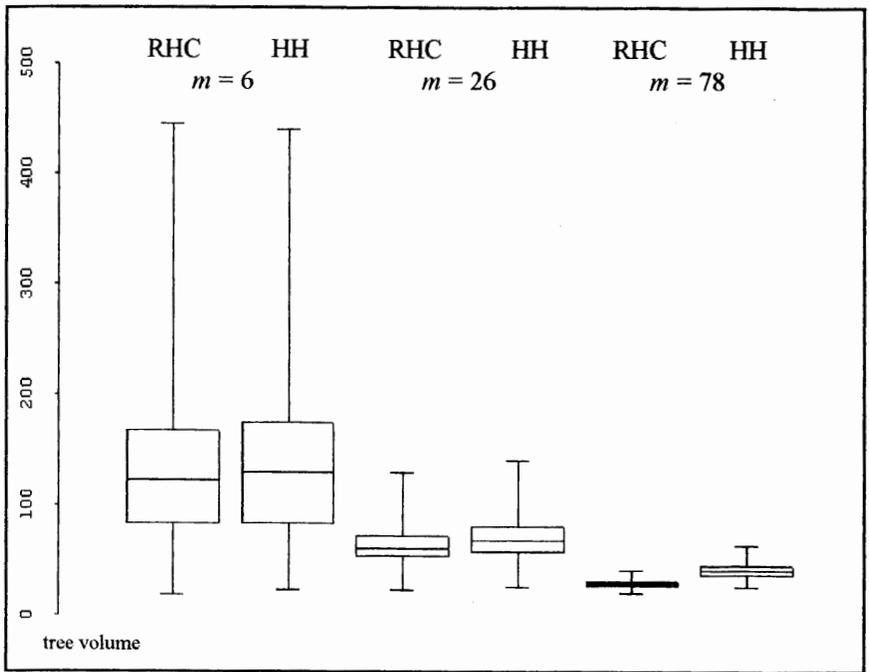


Figure 3.—Boxplots of the Rao-Hartley-Cochran (RHC) and Hansen-Hurwitz (HH) standard error estimates of estimated total tree volume. Sample size is  $m$ .

## REFERENCES

- Chauduri, A. and Mitra, J. 1992. A note on two variance estimators for Rao-Hartley-Cochran estimator. *Communications in Statistics - Theory and Methods*, 21: 3535-3543.
- de Vries, P. G. 1989. *Sampling theory for forest inventory*. New York, NY: Springer-Verlag. 399 p.
- Deshpande, M. N. 1984. A note on Rao, Hartley and Cochran's method. *Sankhyā*, 36: 114-116.
- Hansen, M.H. and Hurwitz, W.N. 1943. On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 20: 426-432.
- Ohlsson, E. 1989. Variance estimation in the Rao-Hartley-Cochran procedure. *Sankhyā*, B, 51: 348-361.
- Rao, J.N.K., Hartley, H.O., and Cochran, W.G. 1962. On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society B*, 24: 482-491.
- Roesch, F. A., Jr., Green, E. J., and Scott, C. T. 1993. An alternative view of forest sampling. *Survey Methodology*, 19: 199-204.
- Singh, R. and Kishmore, L. 1975. On Rao, Hartley and Cochran's method. *Sankhyā*, C, 37:88-94.
- Tierney, L. 1990. *LISP-STAT: An object-oriented environment for statistical computing and dynamic graphics*. New York, NY: John Wiley. 397 p.

## BIOGRAPHICAL SKETCH

Jeffrey S. Pontius is an Assistant Professor in the Department of Statistics and the Center for Applied Statistics, Kansas State University. His research interests are sampling and stochastic processes in ecology.

# Increasing Spatial Precision and Accuracy for Monitoring Peatlands in Switzerland by Remote Sensing Techniques

Michael Köhl<sup>1</sup>, Hans Jörg Schnellbacher<sup>2</sup>, and Andreas Grünig<sup>3</sup>

---

**ABSTRACT.** - In Switzerland about 1700 peatlands are considered to be 'sites of national importance' and are protected on a constitutional level. In 1994 a peatland monitoring program has been initiated. One of the very first steps of this program is to develop a monitoring design that provides reliable and cost-efficient information on the current state and development of national peatlands. A combined photo-/ field sampling design is compared to a field-/ based, one-phase approach in a Monte Carlo simulation study. The results show that concerning variance, bias and root mean square error the combined approach proved superior compared to pure field assessments.

---

## INTRODUCTION

In 1984 and 1990 respectively, the Swiss national bog and fenland inventories were carried out (Grünig et al., 1986, Broggi, 1990). The area of Swiss mire habitats identified by these two inventories totals approximately 30'000 ha, which is less than 1% of the total area of Switzerland. These inventories included approximately 1700 mires with a total of 20000 ha, which are regarded to be 'Sites of National Importance' (Grünig, 1994, Küttel, 1995). As by federal legislation those sites have to be protected and monitored, a concept for monitoring national peatlands to meet the legal requirements had to be developed. The focus of the

---

<sup>1</sup> Project leader, Inventory Methods Research Group, Swiss Federal Institute for Forest, Snow and Landscape Research, 8903 Birmensdorf, Switzerland

<sup>2</sup> Research forester, Inventory Methods Research Group, Swiss Federal Institute for Forest, Snow and Landscape Research, 8903 Birmensdorf, Switzerland

<sup>3</sup> Project leader, Advisory Service for Mire Conservation, Swiss Federal Institute for Forest, Snow and Landscape Research, 8903 Birmensdorf, Switzerland

monitoring program is to obtain results for the entirety of Swiss peatlands and to describe the situation of individual sites (objects).

Long term monitoring of Swiss peatlands has to cover two different objectives: 1) to detect changes in the total area of peatlands and 2) to detect changes of the area covered by distinct vegetation types. As vegetation cover is an indicator for the quality of site conditions, vegetation cover serves to the purpose of monitoring site changes as well. The methods applied to meet the first objective have been described elsewhere (Kleinn, 1991, Köhl and Kleinn, 1994, Köhl, 1994) and are not subject of this paper.

Two monitoring approaches will be presented: surveys based on systematically distributed sample plots, and a two-phase sampling design based on the combination of data from aerial photographs and field assessment. The accuracy (bias) of both methods as well as the precision for estimating current state and change is investigated by simulation studies and pilot surveys on a test site.

### SAMPLING ALTERNATIVES

A traditional approach in vegetation surveys is to establish and assess a set of systematically distributed sampling units and remeasure it at successive occasions. In forest surveys this design is known as control sampling (Schmid, 1967) or continuous forest inventory (Stott, 1947). On each plot either the percentage of area cover by species or the presence or absence of a species can be recorded. Preliminary studies have shown that the assessment of the percentage area cover on sample plots is prone to subjective observer bias, which cannot be reduced by sample size (see Gertner and Köhl, 1992, for details). Thus we will concentrate on the second approach, which is more robust in respect of observer bias.

The proportion of the reference area covered by a species can be estimated by defining a variable  $y$ , which can take the value 1 if the species is observed on the plot and the value 0 in all other cases. For large populations the proportion  $\hat{P}$  can be estimated as follows:

$$\hat{Y} = \frac{\sum y_i}{n} = \frac{a}{n} = \hat{P} \quad [1]$$

with variance

$$\hat{S}^2 = \frac{\sum (y_i - \hat{Y})^2}{n - 1} = \frac{n}{n - 1} \hat{P}(1 - \hat{P}) \quad [2]$$

where  $n$  is the number of sampling points and  $y_i$  is the observation on plot  $i$ . The variance of  $\hat{P}$  is estimated by

$$v(\hat{P}) = \frac{\hat{P}(1 - \hat{P})}{n - 1} \quad [3]$$

If the survey is based on remeasurement of the same set of permanent plots, change between two successive occasions,  $\hat{C}$ , and its variance  $v(\hat{C})$ , can be calculated by taking the difference of the proportions  $\hat{P}_i$  at each occasion.

$$\hat{C} = \hat{Y}_2 - \hat{Y}_1 = \hat{P}_2 - \hat{P}_1 \quad [4]$$

$$v(\hat{C}) = \text{var}(\hat{P}_1) + \text{var}(\hat{P}_2) - 2r_{yx}\sqrt{\text{var}(\hat{P}_1)}\sqrt{\text{var}(\hat{P}_2)} \quad [5]$$

where

$$\hat{Y}_i = \hat{P}_i = \text{proportion at occasion } i \ (i=1,2)$$

$r_{yx}$  = correlation coefficient of observations at occasion 2 and occasion 1

In forest survey applications it has often been shown that cost-efficiency can be improved by combining information assessed in aerial photographs and field assessments. A two-phase (or double) sampling for stratification design has been introduced in the second Swiss national forest inventory (Köhl, 1994), where in a first phase strata weights are assessed on a large sample of aerial photography plots and variables of interest are assessed on a smaller sample of field plots. This design is proposed as an alternative for monitoring the area cover percentage in the Swiss peatland monitoring program.

Applying a double-sampling for stratification design the mean proportion of area cover of a plant species,  $\hat{Y}_{ds}$  and its variance  $v(\hat{Y}_{ds})$  is estimated as follows:

$$\hat{Y}_{ds} = \sum_{h=1}^L \frac{n_h}{n} \hat{Y}_h = \hat{P} \quad [6]$$

$$v(\hat{Y}_{ds}) = \sum_{h=1}^L \frac{n_h - 1}{n - 1} \frac{n_h}{n} v(\hat{Y}_h) + \sum_{h=1}^L \frac{1}{n - 1} \frac{n_h}{n} (\hat{Y}_h - \hat{Y}_{ds})^2 \quad [7]$$

where

$\hat{Y}_h$  = mean value in stratum h, h=1,...L

$v(\hat{Y}_h)$  = variance of  $\hat{Y}_h$ , h=1,...L

$n'_h$  = number of photo plots in stratum h, h=1,...L

$n'$  = total number of photo plots =  $\sum n'_h$

L = number of strata

## DATA

Among the peatlands covered by pilot surveys a mire called 'Gross Moos' in the Schwändital, Canton of Glarus, has been selected to provide basic data on plant communities and their spatial distribution for a simulation study. Aerial photographs (CIR, 1:5000) of the Grossmoos were taken in 1994. The aerial photographs were scanned and analyzed by means of digital photogrammetry using a HELAVA DPW 770 station. The borderline of the 'Gross Moos' was plotted and patches within the total area were assigned and delineated according to differences in texture and color. Similarities of texture and color were used to form nine strata, which can be regarded to reflect distinct plant communities. Thus the construction of strata was not founded on the occurrence of vegetation types or plant communities observed on the ground, but exclusively on attributes (texture and color) assessed in aerial photographs. Each of the delineated patches was grouped into one of the strata and the results of the photo-interpretation were printed in mapped form. Based on a preliminary map a vegetation survey was conducted. For the application of a two-phase design, however, only strata have to be assigned to photo plots, but no map is required. For each patch a field assessment was conducted and the plant species observed and their cover recorded. The cover of each plant species within each patch, p, was assessed in five classes:

- |                 |                 |
|-----------------|-----------------|
| - not present   | - 0% < p ≤ 0.1% |
| - 0.1% < p ≤ 1% | - 1% < p ≤ 10%  |
| - > 10%         |                 |

The results of the vegetation survey were later on used to assign vegetation types to the strata formed by aerial photo-interpretation. For the simulation study six plant species reflecting different spatial distributions and abundance classes were selected. Table 1 summarizes the spatial distribution and the proportion of area cover for the selected species.

**Table 1. - Selected species and corresponding spatial distribution and area cover of patches, in which species have been observed.**

Species	spatial distribution of patches	proportion of area cover within patches
Andromeda polifolia	present on contiguous part of the area	all classes
Anthoxantum odoratum	clustered	frequent, majority >0.1%
Scheuchzeria palustris	clustered in small parts of the area	frequent, majority >0.1%
Senecio alpinus	clustered at the margin	rare, majority < 1%
Tofieldia calyculata	present on small contiguous part of the area	frequent, majority >0.1%
Vaccinium myrtillus	present on majority of patches	frequent, majority >0.1%

## SIMULATION STUDY

In a Monte Carlo simulation study 1-phase and 2-phase sampling designs were tested. The field plots as well as the photo plots were systematically distributed over the entire area. In two phase designs the photo plots were used to estimate the stratum weights. On field plots the presence or absence of each of the six species was recorded. Table 2 summarizes the sampling designs and grid densities included in the simulation study.

**Table 2. - Design alternatives applied in the simulation study**

Sampling design	Grid density	Sample size
1-phase sampling	5 * 5 m	6369
	10 * 10 m	1599
	20 * 20 m	402
2-phase sampling for stratification	1st phase <sup>a</sup> : 5 * 5 m	1st phase <sup>a</sup> : 6369
	2nd phase <sup>b</sup> : 10 * 10 m	2nd phase <sup>b</sup> : 1599
	1st-phase <sup>a</sup> : 10 * 10 m	1st phase <sup>a</sup> : 1599
	2nd phase <sup>b</sup> : 20 * 20 m	2nd phase <sup>b</sup> : 402

<sup>a</sup> aerial photography <sup>b</sup> field assessment

The simulation study was based on the map obtained by the aerial photo-interpretation and the vegetation survey. In the simulation study an exact proportion of area cover within the observed cover class was randomly selected and assigned to each patch. To simulate the spatial distribution within each patch a matrix of pixels with 2.5\*2.5m spatial resolution was laid over the entire area of the 'Gross Moos'. For each pixel a random number was drawn and compared to the proportion of area coverage of a species within the patch, in which the pixel is located. If the random number was smaller than the randomly assigned proportion, the species under concern was assumed to be present on the plot, in all other cases the species was assumed not to be present on the pixel. To test the efficiency of

the 5 design alternatives for the estimation of change, the area cover of each species was altered by 10 percent. Thus the absolute value of changes of percent area cover was higher for frequent species than for species with low percent area cover. Based on the randomly generated maps a second map was produced, in which those pixels were randomly selected, where the presence or absence of species changed. This procedure was repeated for each species. The original stratification with 3 strata was retained for all maps.

New maps for current state and changes were generated for each of the 1000 iterations and the 5 sampling alternatives were simulated in each map. For each sampling alternative, each species, and each iteration the following parameters were estimated:

- current state of percent area cover and its variance using the equations presented above,

- the bias, where  $\text{bias} = |P_{\text{true}} - \hat{P}|$ , and

- the root mean square error RMSE, where  $\text{RMSE} = \sqrt{v(\hat{P}) + \text{bias}^2}$

According to Cochran (1977) the RMSE can be used as a criterion of the accuracy of an estimator.

## RESULTS

In Table 3 to Table 8 the results for the six species are presented. For each of the 5 design alternatives the estimates proportion,  $\hat{P}$ , the bias, the standard error of  $\hat{P}$ ,  $s(\hat{P})$ , and the root mean square error, rmse, are given.

The standard error of ( $\hat{P}$ ) is much smaller for the 2-phase sampling designs than for the 1-phase designs. This holds for both, current values and change, and can be observed independent of the spatial distribution and the frequency of the species. Even for the denser grid size the variance of the 1-phase designs is larger than for the coarsed grids of the 2-phase design. For the estimation of current state and change the bias is smaller for the 2-phase designs with one exception (10\*10/5\*5m grid for *vaccinium myrtillus*). The size of the bias of the 1-phase and 2-phase designs is similar, when the terrestrial grid sizes of the 1-phase designs and the grid size of the photo phase in 2-phase designs are equal. The same effect can be observed for the root mean square error. The studied 2-phase designs arrive at the same root mean square with one quarter of the field plots of a 1-phase design. As the ratio of change was the same in all strata, an improvement is likely for situations, where the changes of cover vary in individual strata.

**Table 3: Scheuchzeria palustris ( $P_{true} = 1.1416$ ,  $C_{true} = 0.1041$ )**

		Grid size [m]				
field:		5*5	10*10	20*20	10*10	20*20
photo:					5*5	10*10
current state	$\hat{P}$	1.2199	1.0175	1.0445	1.0349	1.1521
	bias	0.0782	0.1241	0.0971	0.1067	0.0105
	$s(\hat{P})$	0.1371	0.2493	0.4970	0.0015	0.0023
	rmse	0.1711	0.3157	0.6078	0.1564	0.3104
changes	$\hat{C}$	0.1114	0.0939	0.0970	0.0956	0.1067
	bias	0.0073	0.0101	0.0070	0.0085	0.0027
	$s(\hat{C})$	0.0409	0.0649	0.0883	0.0001	0.0002
	rmse	0.0525	0.1003	0.1768	0.0616	0.1389

**Table 4: Vaccinium myrtillus ( $P_{true} = 8.3311$ ,  $C_{true} = 0.8335$ )**

		Grid size [m]				
Field:		5*5	10*10	20*20	10*10	20*20
Photo:					5*5	10*10
Current state	$\hat{P}$	8.3827	8.2727	8.1731	8.1376	8.1916
	bias	0.0516	0.0584	0.1580	0.1935	0.1395
	$s(\hat{P})$	0.3466	0.6875	1.3625	0.0173	0.0234
	rmse	0.3972	0.8147	1.6503	0.3973	0.7934
Change	$\hat{C}$	0.8393	0.8291	0.8386	0.8153	0.8386
	bias	0.0057	0.0045	0.0050	0.0183	0.0050
	$s(\hat{C})$	0.1139	0.2243	0.4352	0.0022	0.0021
	rmse	0.1442	0.2993	0.5890	0.1707	0.3398

**Table 5: Andromeda polifolia ( $P_{true} = 0.5267$ ,  $C_{true} = 0.0501$ )**

		Grid size [m]				
Field:		5*5	10*10	20*20	10*10	20*20
Photo:					5*5	10*10
Current state	$\hat{P}$	0.5353	0.5353	0.7007	0.5348	0.6274
	bias	0.0086	0.0085	0.1740	0.0081	0.1007
	$s(\hat{P})$	0.0908	0.1795	0.3975	0.0045	0.0012
	rmse	0.1077	0.2237	0.5165	0.1145	0.2494
Change	$\hat{C}$	0.0496	0.0470	0.0622	0.0469	0.0564
	bias	0.0005	0.0031	0.0121	0.0031	0.0063
	$s(\hat{C})$	0.0263	0.0382	0.0577	0.0002	0.0001
	rmse	0.0343	0.0671	0.1147	0.0425	0.0821

**Table 6: *Tofieldia calyculata* ( $P_{\text{true}} = 0.0962, C_{\text{true}} = 0.0094$ )**

		Grid size [m]				
Field:		5*5	10*10	20*20	10*10	20*20
Photo:					5*5	10*10
Current state	$\hat{P}$	0.0954	0.1094	0.1418	0.1073	0.1482
	bias	0.0008	0.0133	0.0456	0.0112	0.0521
	$s(\hat{P})$	0.0378	0.0720	0.1214	0.0007	0.0012
	rmse	0.0491	0.1059	0.2116	0.0623	0.1583
Change	$\hat{C}$	0.0097	0.0124	0.0172	0.0122	0.0180
	bias	0.0003	0.0031	0.0078	0.0028	0.0087
	$s(\hat{C})$	0.0082	0.0117	0.0170	0.0001	0.0002
	rmse	0.0134	0.0227	0.0320	0.0170	0.0256

**Table 7: *Anthoxantum odoratum* ( $P_{\text{true}} = 3.0541, C_{\text{true}} = -0.3047$ )**

		Grid size [m]				
Field:		5*5	10*10	20*20	10*10	20*20
Photo:					5*5	10*10
Current state	$\hat{P}$	3.0558	3.0882	3.7194	3.0964	3.7019
	bias	0.0017	0.0341	0.6653	0.0423	0.6479
	$s(\hat{P})$	0.2151	0.4312	0.9394	0.0038	0.0081
	rmse	0.2470	0.5096	1.2524	0.2313	0.7371
Change	$\hat{C}$	0.3082	0.3091	0.3776	0.3098	0.3756
	bias	0.0035	0.0044	0.0729	0.0051	0.0709
	$s(\hat{C})$	0.0689	0.1352	0.2646	0.0004	0.0007
	rmse	0.0878	0.1779	0.3856	0.0986	0.2282

**Table 8: *Senecio alpinus* ( $P_{\text{true}} = 1.9300, C_{\text{true}} = 0.1767$ )**

		Grid size [m]				
Field:		5*5	10*10	20*20	10*10	20*20
Photo:					5*5	10*10
Current state	$\hat{P}$	1.9203	1.9548	2.1542	1.9571	2.2573
	bias	0.0097	0.0248	0.2242	0.0271	0.3273
	$s(\hat{P})$	0.1717	0.3451	0.7191	0.0019	0.0045
	rmse	0.1947	0.4011	0.8690	0.1730	0.4832
Change	$\hat{C}$	0.1767	0.1828	0.2119	0.1830	0.2227
	bias	0.0000	0.0061	0.0352	0.0063	0.0459
	$s(\hat{C})$	0.0520	0.1009	0.1713	0.0002	0.0004
	rmse	0.0661	0.1374	0.2818	0.0788	0.1891

## CONCLUSION

The results of the simulation study showed that the same root mean square error can be achieved by 2-phase sampling designs, which use a large set of photo plots to estimate strata sizes and a subsample of field plots, with about  $\frac{1}{4}$  of the field

plots of a comparable 1-phase design. The efficiency of a 2-phase design proved to be superior.

However, the optimal design can only be selected, when cost-efficiency is concerned. The optimal design among the design alternative is the one, which either leads to the smallest variance for given cost or which results in the smallest cost for a desired variance. For a desired variance the total cost of obtaining aerial photographs, aerial photo interpretation and amortization of photo interpretation equipment should be less than 75 percent of the total cost of a strictly field based assessment, referring to the results obtained in the simulation study. A 2-phase design would be superior in terms of cost-efficiency, if the cost-ratio of field and photo plots would be 1.33:1. Experience has shown that the cost ratio is much better referring to photo samples. Taking into account other side effects of aerial photography such as the documentation of a historical situation, providing a piece of evidence for later investigations, and allowing retrospective analysis at later points in time, the superiority of the 2-phase approach is evident.

In many situations vegetation surveys are purely based on field assessments. The data recorded for a single plot are extrapolated to the area represented by the plot, i.e. the surrounding area, and used to produce pixel maps. The grid density of 1-phase sampling designs has to be fairly large to give a realistic situation of the spatial distribution and frequency of species in the area under concern. Especially, the necessary sample size to detect sensitive (and meaningful) changes of spatial patterns should not be underestimated. A stratification in aerial photographs provides a more realistic picture of the spatial distribution of vegetation types. If a 2-phase design is applied, cost efficiency for the estimation of current state and changes is better compared to 1-phase designs. Thus, whenever possible, the combination of field assessments and photo interpretation should be given preference.

## ACKNOWLEDGMENT

We would like to thank Dr. Bernhard Oester, WSL, for providing the results of the photo-interpretation. Dr. Risto Päivinen, European Forest Institute (EFI), and Dr. Peter Brassel, WSL, reviewed the manuscript and gave helpful comments.

## REFERENCES

- Broggi, M.F. (Ed.), 1990: *Inventar der Flachmoore von nationaler Bedeutung. Entwurf für die Vernehmlassung*. Bern, Bundesamt für Umwelt, Wald und Landschaft (BUWAL) 79 p.
- Cochran, W.G. 1977. *Sampling techniques*, J. Wiley & Sons, New York. 428 p.
- Gertner, G.Z., Köhl, M. 1992. An assessment of some nonsampling errors in a national survey using an error budget. *Forest Science*. Vol. 38, No. 3. 525-538

- Grünig, A.; Vetterli, L.; Wildi, O., 1986: Die Hoch- und Uebergangsmoore der Schweiz - eine Inventarwertung. Ber. Eidgenöss. Forstinst. Wald Schnee Landsch. 281, 62 p.
- Grünig, A. (Ed.), 1994: Mires and Man. Mire conservation in a densely populated country - the Swiss experience. Excursion guide and symposium proceedings of the 5th field symposium of the International Mire Conservation Group (IMCG) to Switzerland 1992. Birmensdorf, Swiss Federal Institute for Forest, Snow and Landscape Research. 415 p.
- Kish, L. 1965. Survey sampling, J. Wiley & Sons, New York, 643 p.
- Kleinn, C. 1991. Der Fehler von Flächenschätzungen mit Punktrastern und linienförmigen Stichproben, Universität Freiburg., Mitt. Abt. Forstl. Biom., Nr. 91-1
- Köhl, M. 1994. Statistisches Design für das zweite Schweizerische Landesforstinventar: Ein Folgeinventurkonzept unter Verwendung von Luftbildern und terrestrischen Aufnahmen, Mitteilungen der Eidgenössischen Forschungsanstalt für Wald, Schnee und Landschaft, Birmensdorf, Band 69, Heft 1, 141 S.
- Köhl, M., Kleinn, C. 1994. Stichprobeninventuren zur Waldflächenschätzung. Allgemeine Forst- und Jagdzeitung, 165 (12): 229-231
- Küttel, M., 1995: Moorschutz in der Schweiz - Stand und Ziele. *Telma* 25. 177-192.
- Schmid, P. 1967. Die Weiterentwicklung der Leistungskontrolle in der Schweiz. *Wiss.Z.Tech. Univ. Dresden*, Jg. 16/2
- Stott, C.B. 1947. Permanent Growth and Monitoring Plots in Half the Time. *J.for.* 37: 669-673

## **BIOGRAPHICAL SKETCH**

Michael Köhl is project leader and head of the Inventory Methods Research Group, Swiss Federal Institute for Forest, Snow and Landscape Research (WSL). He graduated from University of Freiburg, Germany, in 1983 and holds a diploma in forestry and a PhD in biometry (1986). He is lecturer at the Swiss Federal Institute of Technology (ETH), Zurich, where he habilitated in 1993. Michael is responsible for the statistical and remote sensing methods of the Swiss National Forest Inventory and is leader of IUFRO S4.11 (Statistics, Mathematics and Computers).

Hans Jörg Schnellbacher is a research forester with the Inventory Methods Research Group, WSL. He graduated from University of Freiburg, Germany, in 1992 and holds a diploma in forestry. Hans Jörg provides statistical consulting for the Swiss National Forest Inventory and is responsible for the data-base management and the statistical analysis of the inventory data.

Andreas Grünig is project leader and head of the Advisory Service for Mire Conservation, WSL. He graduated at the Swiss Federal Institute for Technology (ETH), Zürich and holds a diploma as a scientist (biologist). Andreas is responsible for the Swiss Mire Monitoring Programme.

# Spatial and Temporal Models in Contextual Classification

Bo Ranneby<sup>1</sup>

---

**Abstract.** Different kinds of spatial and temporal models and their use for contextual classification of satellite images are discussed. Models describing the ground cover and its changes over time and models of the spectral signatures in multitemporal images are studied. Isotropic autocorrelation models as well as mosaic models are used for quantification of the variation in space and time.

For the different models parameter estimation, robustness properties, and classification performance are examined. The use of appropriate estimation techniques is crucial for the result of the classification. The use of multitemporal images makes it easier to construct robust estimation methods, and also improves the classification results.

---

## INTRODUCTION

The purpose of this paper is to discuss some models and estimation problems which appear in the classification of remote sensing images. The true image on a two-dimensional regular lattice  $D$  will be denoted by  $\{C(s), s \in D\}$ . The measurement of pixel  $s$  is given by  $Z(s)$ . Unfortunately, there is no one to one correspondence between the  $C$ - and  $Z$ -process, but they are contaminated with noise. The classification algorithms referred to in the present paper relies on Bayesian methods. Let  $\pi(c)$  denote the prior distribution of the  $C$ -process and let  $f(z|C)$  denote the conditional distribution of  $Z$  given  $C$ . Then the posterior distribution of  $C$  given  $Z$  is  $p(C|Z) \propto \pi(C) f(Z|C)$ . The maximum a posteriori estimator is the one that maximizes  $p(C|Z)$ . The crucial steps in the derivation of the classification algorithm are modelling the different processes.

## MOSAIC-MODELS

For many of the applications we have in mind mosaic-models or models related to mosaic models will give a realistic description of true classes of the image and the development over time of these classes. As a starting point to the

---

<sup>1</sup> Department of Forest Resource Management and Geomatics  
Swedish University of Agricultural Sciences  
S-901 83 UMEÅ  
Sweden

study of mosaic-models we consider stochastic methods of dividing the plane into contiguous subsets called cells.

A simple (to handle mathematically) mosaic is the following line-mosaic, see Pielou (1964) and Switzer (1965). The plane is partitioned into convex cells by locating straight lines at random according to the following procedure. Choose points  $(\alpha, \theta)$ , in the infinite strip  $0 \leq \theta < \pi$ , according to a Poisson process with intensity  $\gamma$  points per area unit and associate with each point  $(\alpha, \theta)$  the line

$$y \cos \theta + x \sin \theta - \alpha = 0$$

Each cell is then independently assigned its color according to a probability law, which selects class  $i$  with probability  $p_i$ . If there is only two colors to choose between we get the Pielou-Switzer model.

This model has also been used by Owen (1984) in contextual classification. If the intensity parameter of the Poisson field is small relative to the size of the pixels then:

(i) only three types of patterns X, L, and T are possible in the first order neighborhoods (s, Ns, Ss, Es, Ws).

X pattern	T pattern	L pattern
$i$	$j$	$j$
$i \quad i \quad i$	$i \quad i \quad i$	$j \quad i \quad i$
$i$	$i$	$i$

All rotations are assumed to be equally likely.

(ii) each neighborhood has the same probability of intersecting one boundary line, and small probability of intersecting two or more boundaries.

(iii) the portion of a boundary within a neighborhood is a line segment, so that a single boundary can induce only an L- or a T-pattern.

To this model the variation in time is naturally incorporated by independently assigning a stochastic process  $\xi(t)$  to each cell. The processes give the color of the cells at different times.

As the size of forest stands is large compared to the pixel size this kind of models will be useful in contextual classification of forested areas.

Let us now consider another way of generating the cell partition. Consider points (centers) located in the plane according to a Poisson process with density  $\lambda$ . Attach to each cell its Dirichlet cell consisting of those points that are closer to this center than to any other. The Dirichlet cells of a realization of a Poisson process constitute a random division of the plane into polygons varying in size and shape. The variance of the area of a cell is  $0.28018/\lambda^2$ , see Matérn (1979). The mosaic, and its variation over time, is then obtained by independently assigning a stochastic process to each cell giving the color of the cell at different times.

For space-time models of this kind it holds that the correlation in space and time is the product of the spatial correlation and the correlation in time given by the correlation of the process  $\xi(t)$ . As shown in Ranney (1982) this holds for a rather large class of models. If we have a partition into cells so the cell-model is isotropic and we to each cell, independently of the cell partition and of other cells, have attached stochastic processes  $\xi_1(t), \xi_2(t), \dots$  to the cells, then the correlation function of the variation in space and time is given by the product of the correlation function of the variation in space and the correlation function of the variation in time. It is crucial that the  $\xi$ -processes have the same distribution. The result can be extended to models where neighboring cells may be dependent.

The mosaic models above describe the variation on the ground. An extremely simple model describing probabilities of different class configuration are obtained by assuming conditional independence. The conditional probability of getting class configurations a,b,c,d given class k in the centre pixel is given by:

$$g(a,b,c,d | k) = p(a | k) p(b | k) p(c | k) p(d | k)$$

This model has been used by Haslett (1985) for contextual classification. It is easily extended to images from several occasions, see Flygare (1993).

### MODELS FOR SPECTRAL REFLECTANCES

For modelling the spectral reflectances autocorrelation models have shown to be useful. In most cases these models are assumed to be multivariate Gaussian processes. The vector  $X(s)$  of observed reflectances for pixel  $s$  at time  $t$  is assumed to be a sum of two independent variables  $Y(s)$  and  $\varepsilon(s)$ . The  $\varepsilon$ -process is autocorrelated noise, while the  $Y$ -process gives the true signal. Usually we have images from one or two occasions. When we have two images  $t$  equals 1 or 2. When we have only one image the index  $t$  is omitted. The following assumptions will be made for the autocorrelation models.

- A<sub>1</sub>: The random fields  $\{Y(s), s \in D\}$  and  $\{\varepsilon(s), s \in D\}$  are independent
- A<sub>2</sub>: The random field  $\{\varepsilon(s), s \in D\}$  is Gaussian with expectation zero and  $\text{Cov}(\varepsilon(s), \varepsilon^T(s_1)) = \rho^d \theta \Sigma$ , where  $d = |s - s_1|$  is the distance between  $s$  and  $s_1$ .
- A<sub>3</sub>: The random field  $\{Y(s), s \in D\}$  is Gaussian with expectation zero and covariance matrix  $(1 - \theta)\Sigma$  and  $\text{Cov}(Y(s), Y^T(s_1)) = 0, \quad s \neq s_1$

The parameter  $\theta, 0 \leq \theta \leq 1$ , is a proportion parameter telling us the amount of variation in  $X(s)$  coming from  $\varepsilon(s)$  vs  $Y(s)$  and  $\rho, 0 \leq \rho < 1$ , is the spatial

autocorrelation between noise vectors  $\varepsilon(s)$  and  $\varepsilon(s_1)$  for pixels a unit distance apart.

The assumptions  $A_1$ - $A_3$  are enough for one image models. For two image models the following additional assumptions are made:

- $A_4$ : The three random fields  $\{(Y(s,2), Y(s,1)), s \in D\}$ ,  $\{\varepsilon(s,2), s \in D\}$  and  $\{\varepsilon(s,1), s \in D\}$  are independent.
- $A_5$ : The random fields  $\{\varepsilon(s,2), s \in D\}$  and  $\{\varepsilon(s,1), s \in D\}$  are Gaussian with expectation zero and with the same covariance matrix  $\theta\Sigma$ .
- $A_6$ :  $(Y(s,2), Y(s,1))$  and  $(Y(s_1,2), Y(s_1,1))$  are independent for  $s \neq s_1$
- $A_7$ :  $(Y(s,2), Y(s,1))$  have Gaussian distributions with zero expectation and covariance matrix  $(1 - \theta)\Sigma$  for  $(Y(s,2)$  and  $Y(s,1)$  and with cross covariance function  $\text{Cov}(Y(s,2), Y^T(s,1)) = \tau(1 - \theta)\Sigma$ .

To make these models operational the different parameters have to be estimated.

### ESTIMATION PROBLEMS

There are interesting estimation problems associated with several of the models described above. Usually the training sets for the different classes are selected by subjective methods - areas which are typical for the different classes are chosen. This implies that the means may be unbiasedly estimated but the variation within the classes may be heavily underestimated. If so, too many pixels will not be "typical" for any class, which may result in an unnecessary large "doubt"-class. If the training sets are chosen by objective methods this problem is avoided.

For good performance of classification methods based on the autocorrelation model it is necessary to have almost unbiased estimates of the parameters. If the noise-process had been observable the autocorrelation parameters had been estimated by replacing the theoretical moments by their empirical correspondence. Since we can only observe the X-process there is a risk that the autocorrelations in the noise-process are contaminated with the autocorrelations for the true spectral reflectances. To avoid this homogeneous pixelcrosses (the pixel and its N-, S-, W-, and E-neighbors) should be used. If the spectral reflectances are the same for all pixels, then as we want, it is only the noise-process, which causes the autocorrelations. This is also why Hjort et al (1984) recommends homogeneous pixel crosses. This is the best we can do if only one image is available.

If images from two occasions, close in time, are available it is possible to base our estimates on the difference of the two images. Taking the differences over areas with slow changes we get under reasonable assumptions, the difference between the

autocorrelated noise for the two images. This has been studied in Flygare (1996), where she discusses some robustness properties and shows strong consistency of the autocorrelation estimates.

Let us have a closer look at the robustness properties for the two methods. Suppose that assumption  $A_3$  in the previous section is not satisfied but that  $\text{Cov}(Y(s), Y(s_1)) = \eta(d)(1 - \theta)\Sigma$ , where  $\eta(d)$  is an unknown positive correlation function with  $\eta(0) = 1$ . If  $A_3$  holds then  $\eta(d) = 0$  for  $s \neq s_1$ .

For the one image method we get that

$$E \left[ \frac{1}{n_k} \sum (X^k(s) - \mu(k))(X^k(s_1) - \mu(k))^T \right] \rightarrow \eta(d)(1 - \theta)\Sigma + \rho^d \theta \Sigma,$$

As  $\eta(d)$  in most applications is larger than zero this results in an overestimation.

For the difference method the situation is somewhat different. Here we get the following expected value for our estimate:

$$2(1 - \tau)\eta(d)(1 - \theta)\Sigma + 2\rho^d \theta \Sigma.$$

If  $\tau$  is close to 1 we see that the expected value is approximately  $2\rho^d \theta \Sigma$ , which is what we get when all assumptions are fulfilled. As a consequence the difference method is less sensitive to errors in the model.

## MIXTURES OF MULTINORMAL DISTRIBUTIONS

In Taxt et al (1991) mixtures of normal distributions are used for classification of images. They have rather frequently observed bimodal and multimodal empirical distributions for feature vectors from the same class, when working with remote sensing images. If the training sets are objectively selected there are reasons to believe that empirical distributions of this kind will be more common. Taxt et al (1991) used linear combinations of two multinormal distributions to model the distributions for the different classes. A typical model for the density function  $f(x)$  of the feature vector is given by

$$f(x) = (1 - \pi)N_d(\mu_1, K)(x) + \pi N_d(\mu_2, M)(x)$$

The estimation of the parameters in such distributions is known to be a difficult problem, both from statistical and algorithmic points of view.

Example. Let  $\xi_1, \xi_2, \dots$  be i.i.d. observations from a mixture of two bivariate normal distributions. The density function  $f(x, y, \theta)$  is given by

$$f(x, y, \theta) = p h(x, y, \mu_1, \mu_2, \sigma_1, \sigma_2, \rho_1) + (1-p) h(x, y, \mu_3, \mu_4, \sigma_3, \sigma_4, \rho_2),$$

where  $h$  is the density function for a bivariate normal distribution with parameters indicated by the notation. Say that we have observations  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ . If we put  $\mu_1 = x_1, \mu_2 = y_1$  and let  $\sigma_1, (\text{or } \sigma_2)$  go to zero, then the likelihood function tends to infinity. Consequently, the ML-method is not suitable.

The same problem appears also for univariate distributions. In Ranney (1984) a new estimation method, the *Maximum Spacing (MSP) Method*, was introduced which gives consistent estimators also for distributions of this kind. The method is closely related to the Maximum Likelihood Method - both can be derived as approximations of the Kullback-Leibler information. The MSP-method is constructed so that each contribution is bounded from above for any parameter combination.

The same estimation method has been used by Cheng & Amin (1983) for three-parameter Weibull and gamma models. The method is also discussed in Titterton (1985), where he states that "in principle, of course it would be possible to treat multivariate data by grouping: although definition of the multinomial cells would be more awkward."

Since we do not have any natural order relation in  $R^d$  when  $d > 1$ , it is not at all obvious how the generalization shall be carried out. In the univariate case we attach to each observation  $\xi_i$  the variable  $\eta_i(n) = (n + 1)$  "distance to the nearest observation to the right of  $\xi_i$ ".

In  $R^d, d > 1$ , we cannot use  $\eta_i(n)$ . However, to each observation,  $\xi_i$ , we can attach its Dirichlet cell. The Dirichlet cell  $D_n(\xi_i)$  surrounding  $\xi_i$  consist of all points  $x \in R^d$  which are closer to  $\xi_i$  than to any other observation.

Let  $\xi_1, \xi_2, \dots$  be i.i.d. random vectors in  $R^d$  with true density function  $g(x)$  and suppose we assign a model with density functions  $\{f(x, \theta); \theta \in \Theta\}$ , where  $\Theta \subset R^d$ . Let  $A$  be a measurable subset of  $R^d$ . Then  $P_\theta(A)$  denotes the probability that  $\xi_i$  belongs to the set  $A$  calculated under the assumption that the density function is given by  $f(x, \theta)$  (i.e.  $P_\theta(A) = \int_A f(x, \theta) dx$ ). The true probability distribution is denoted by  $P_0(\cdot)$ .

The Dirichlet cells  $D_n(\xi_i), i = 1, 2, \dots, n$  split  $R^d$  into  $n$  identically distributed

random sets. Since  $\sum_{i=1}^n P_0(D_n(\xi_i)) = 1$ , symmetry arguments give

$$E[P_0(D_n(\xi_i))] = \frac{1}{n} \quad \text{or} \quad P_0(D_n(\xi_1)) \cong \frac{1}{n}. \quad \text{Since it also holds that}$$

$$P_0(D_n(\xi_i)) \cong |D_n(\xi_i)|g(\xi_i),$$

we get that  $g(\xi_i) \cong 1/n |D_n(\xi_i)|$ . Here  $|D_n(\xi_i)|$  denotes the "volume of the set  $D_n(\xi_i)$ ".

Consequently, the random variable  $n|D_n(\xi_i)|$  may be interpreted as a non-parametric estimate of  $1/g(\xi_i)$ . Since

$$nP_\theta(D_n(\xi_i)) \cong n|D_n(\xi_i)|f_\theta(\xi_i) \cong f_\theta(\xi_i)/g(\xi_i),$$

it is intuitively obvious that

$$\frac{1}{n} \sum_{i=1}^n \ln nP_\theta(D_n(\xi_i)) \tag{1}$$

will be an approximation of the Kullback-Leibler distance,  $I(g, f_\theta)$ , defined by

$$I(g, f_\theta) = \int g(x)(\ln g(x) - \ln f_\theta(x))dx.$$

The approximation above is analogous to the univariate approximation given in Ranney (1984) and the *Maximum Spacing Estimate* of  $\theta$  is obtained by maximizing approximation (1) above.

For each class we estimate the mixture proportion, the mean vectors and the covariance matrices. Remote sensing images have the property that the interclass distances are rather small so the different parts of the mixture distributions are not easy to separate. This makes it difficult to find good starting values which are crucial for other estimation methods. Although the multivariate MSP- method theoretically give an attractive estimation method much work remains until all practical problems are solved.

Taxt et al (1991) found a substantial increase in the correct classification rates when using mixture models instead of multivariate normal distributions.

## CLASSIFICATION RESULTS

In Flygare (1995) the performance of some contextual classification methods is evaluated on Landsat TM data. Haslett's model was used for the conditional class distribution and in three of the methods she used models with autocorrelated spectral reflectances. Methods utilizing images from one and two occasions were compared. There was no surprise that an autocorrelation model utilizing information from two occasions gave the best classification results. Mean Absolute Deviation was used to compare the classification results with the evaluation data. A 21% improvement was obtained with the two image autocorrelation method compared to the best one image method.

## BIOGRAPHICAL SKETCH

Bo Ranneby is professor at the Swedish University of Agricultural Sciences. He took a Ph.D. in mathematical statistics at Umeå University in 1975. In 1992 he was elected as an ordinary member of the International Statistical Institute. He is head of the Department of Forest Resource Management and Geomatics.

## REFERENCES

- Cheng, R.C.H. & Amin, N.A.K. (1983). Estimating parameters in continuous univariate distributions with a shifted origin. *J.R. Statist. Soc. B.* 45, 394-403.
- Flygare, A-M. (1993). Contextual classification using multi-temporal Landsat TM data. Statistical Research Report, University of Umeå.
- Flygare, A-M. (1995). Comparing some contextual classification methods using Landsat TM. Research report no 7, Institute of Mathematical Statistics, Umeå University.
- Flygare, A-M. (1996). An alternative estimation method for spatial autocorrelation parameters. Institute of Mathematical Statistics, Umeå University.
- Haslett, J. (1985). Maximum likelihood discriminant analysis on the plane using a Markovian model of spatial context. *Pattern Recognition*, 18, 287-296.
- Hjort, N.L. and Mohn, E. (1984). A comparison of some contextual methods in remote sensing classification. In: *Proceedings of the Eighteenth International Symposium on Remote Sensing of the Environment*. Centre National d'Etudes Spatiales, Paris, 1693-1702.
- Matérn, B. (1979). The analysis of ecological maps as mosaics. In: *Spatial and Temporal Analysis in Ecology*, edited by R.M. Cormack & J.K. Ord, *Statistical Ecology Series*, no 8, 271-288.
- Owen, A. (1984). A neighborhood-based classifier for Landsat data. *Canadian Journal of Statistics*, 12, 191-200.
- Pielou, E.C. (1964). The spatial pattern of two-phase patchworks of vegetation, *Biometrics*, 20, 156-167.
- Ranneby, B. (1982). Stochastic models of variation in time and space. In: *Statistics in Theory and Practice. Essays in Honour of Bertil Matérn*, B. Ranneby, ed. Swedish University of Agricultural Sciences, Umeå, Sweden, 227-245.
- Ranneby, B. (1984). The maximum spacing method. An estimation method related to the maximum likelihood method. *Scand. J. Statist.* 11, 93-112.
- Taxt, T., Hjort, N.L. & Eikvil, L. (1991). Statistical classification using a linear mixture of two multinormal probability densities. Statistical Research Report, no 5, Institute of Mathematics, University of Oslo.
- Titterton, D.M. (1985). Comments on "Estimating parameters in continuous univariate distributions with a shifted origin". *J.R. Statist. Soc. B* 47, 115-116.
- Switzer, P. (1965). A random set process in the plane with a Markovian property, *Ann. Math. Stat.*, 36, 1859-1863.

# Generalized Linear Mixed Models for Analyzing Error in a Satellite-based Vegetation Map of Utah

Gretchen G. Moisen<sup>1</sup>, D. Richard Cutler<sup>2</sup>,  
and Thomas C. Edwards, Jr.<sup>3</sup>

---

**Abstract.**—With the increasing demand for broad-scale vegetation maps for ecosystem management and conservation planning comes the need for flexible tools to assess thematic accuracy of these maps. In this paper, we use a generalized linear mixed model (GLMM) to explore the relationship between thematic accuracy in the blackbrush cover-type of a satellite-based vegetation map of Utah and various topographical and heterogeneity components of that map. Because of the difficulty in accessing many rugged areas of this State, two strata were defined based on proximity to roads. Vegetation type was recorded on heterogeneous linear clusters of sample points within the "off-road" strata, and on randomly distributed sample points within the "road" strata on selected USGS quadrangle maps. A binary response (correctly classified / incorrectly classified) was modeled as a function of both fixed and random effects accounting for spatially autocorrelated observations and different covariance structures for the random effects. The modeling exercise suggested a strong relationship between map error in the blackbrush cover-type of the Colorado Plateau of Utah, and stratum, slope and local heterogeneity.

---

## INTRODUCTION

Thematic accuracy of vegetation cover maps derived from satellite imagery may be related to many factors, including elevation, aspect, slope, local heterogeneity and distance to vegetation boundaries. Exploring the relationship between the components of the vegetation classification model and its uncertainty is a logical step in an analysis of map error that is sensitive to both map use and subsequent improvements of the map.

Although many new techniques are being explored to address map uncertainty, generalized linear mixed models (GLMM's) have yet to be applied. Through a GLMM, data from any one of a variety of continuous and discrete distributions can be linked to a linear structure that may contain both fixed and random effects. GLMMs can also account for correlation among observations as well as among random effects terms in the linear structure (Wolfinger and O'Connell 1993).

<sup>1</sup> USDA Forest Service, Intermountain Research Station, Ogden, UT.

<sup>2</sup> Department of Mathematics and Statistics, Utah State University, Logan, UT.

<sup>3</sup> USDI National Biological Service, Utah Cooperative Fish and Wildlife Research Unit, Utah State University, Logan, UT.

This flexibility may prove valuable in addressing map uncertainty as well as have numerous other broad-scale applications.

In this study, we use a GLMM to explore the relationship between the error in the blackbrush cover-type of a vegetation map of Utah and various topographical and heterogeneity components of that map.

## DATA

A cover-map of Utah, ~219,000 km<sup>2</sup> in size, was developed from a state-wide Landsat Thematic Mapper (TM) mosaic created from 24 scenes at 30 m resolution (Homer et al. in press). A total of 38 cover-types were modeled. Modeling was accomplished using a four step modeling approach. Steps included: (1) the creation of a statewide seamless mosaic of TM images; (2) the subsetting of the mosaic into 3 ecoregions, the Basin and Range, Wasatch-Uinta and Colorado Plateau (after Omernik 1987); (3) the association of 1,758 state-wide field training sites to spectral classes; and (4) the use of ecological parameters based on elevation, slope, aspect and location to further refine spectral classes representing multiple cover-types .

Following development of this cover-map, field data were collected to assess its thematic accuracy (see Moisen et al. 1994 for design considerations, Edwards et al. 1996 for analysis). Of primary interest were estimates of by-class and by-ecoregion accuracy of the map at the base model of one ha. A total of 100 7.5-min quadrangles were randomly selected roughly proportional to the area of the three ecoregions. Two strata were identified on each quadrangle based on proximity to roads. The "road" stratum consisted of all land within 1 km of a secondary or better road. All other lands fell within the "off-road" stratum. On each quadrangle, ten points were randomly selected within the road stratum, while ten were collected in a randomly oriented heterogeneous linear cluster within the off-road stratum. Data were then used to assess map accuracy based on procedures outlined in Edwards et al. (1996).

For this study, a subset of the state-wide data comprised of the blackbrush cover-type within the Colorado Plateau was modeled under a GLMM. Data consisted of 96 sample points collected on two strata (road, off-road) on 15 quadrangles. Anywhere from one and ten blackbrush points were available for each quad/stratum combination. Clustered blackbrush data in the off-road stratum were not necessarily adjacent sample points because blackbrush polygons were often intermixed with other cover-types not considered in this analysis.

## MODEL

Using a logit link function, the binary response (correctly classified / incorrectly classified) was modeled as a function of both fixed and random effects while accounting for several covariance structures for random effects and for spatially autocorrelated errors. For this analysis, the observations on the 96 sample points were coded as 1 when the mapped cover-type agreed with the ground

cover-type, and as 0 when they did not agree. Define  $\mathbf{y}$  to be our data vector of 96 0s and 1s satisfying

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}. \quad (1)$$

We used a logit link function

$$g(\boldsymbol{\mu}) = \log\{\boldsymbol{\mu}/(1 - \boldsymbol{\mu})\} \quad (2)$$

and modeled

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}. \quad (3)$$

Here,  $\boldsymbol{\beta}$  is a vector of unknown fixed effects with known model matrix  $\mathbf{X}$ , and  $\mathbf{v}$  is a vector of unknown random effects with known model matrix  $\mathbf{Z}$ . Assume  $E(\mathbf{v}) = \mathbf{0}$  and  $\text{cov}(\mathbf{v}) = \mathbf{G}$ , where  $\mathbf{G}$  is unknown. An effect may be considered fixed if the inference space is limited to the observed levels of that effect. An effect may be considered random if the inference space is applied to a population of levels, not all of which are observed. Fixed effects considered in this application included both discrete and continuous variables. Because quadrangle maps were randomly selected for subsampling from a population of quadrangles, quadrangles were modeled as random effects. Also,  $\boldsymbol{\varepsilon}$  is a vector of unobserved errors with  $E(\boldsymbol{\varepsilon}|\boldsymbol{\mu}) = \mathbf{0}$  and

$$\text{cov}(\boldsymbol{\varepsilon}|\boldsymbol{\mu}) = \mathbf{R}_{\boldsymbol{\mu}}^{1/2} \mathbf{R} \mathbf{R}_{\boldsymbol{\mu}}^{1/2}. \quad (4)$$

Here  $\mathbf{R}_{\boldsymbol{\mu}}$  is a diagonal matrix containing evaluations at  $\boldsymbol{\mu}$  of the variance function

$$V(\boldsymbol{\mu}) = \boldsymbol{\mu}(1 - \boldsymbol{\mu}). \quad (5)$$

$\mathbf{R}$  and  $\mathbf{G}$  were modeled using covariance structures detailed below.

### Fixed Effects

Nine fixed effects variables were considered. Three were topographical variables used in the classification model itself. These were extracted from a 90 m Digital Elevation Model and include elevation in meters (ELEV), slope in degrees (SLOPE), and aspect. A transformation of aspect (TRASP), used by Roberts and Cooper (1989), takes the form

$$\text{TRASP} = \frac{1 - \cos(\text{aspect} - 30)}{2}. \quad (6)$$

This transformation assigns the highest values to land oriented in a north-northeast direction, the coolest and wettest orientation in Utah.

In addition to the three topographical variables, we considered four different measures of heterogeneity surrounding the sample point. Richness (RICH) is defined as the number of cover-types found in the surrounding 8 pixels. The other three heterogeneity variables, evenness (EVEN) and two measures of diversity ( $D_1$  and  $D_2$ ), are defined in table 1. Higher values for all indices indicate increasing heterogeneity.

**Table 1.—Measures of heterogeneity. Here S equals richness, n equals 8 pixels, and  $n_i$  is the number of pixels belonging to the  $i$ th of S cover-types.**

Variable	Source	Formula
$D_1$	Hill (1973)	$D_1 = \exp\left(-\sum_{i=1}^S \left[\left(\frac{n_i}{n}\right) \ln\left(\frac{n_i}{n}\right)\right]\right) \quad (7)$
$D_2$	Simpson (1949)	$D_2 = \left[\sum_{i=1}^S \frac{n_i(n_i-1)}{n(n-1)}\right]^{-1} \quad (8)$
EVEN	Ludwig and Reynolds (1988)	$\text{EVEN} = (D_2 - 1) / (D_1 - 1) \quad (9)$

Two other fixed effects considered were the minimum distance in meters to a different map cover-type (DIST) and a variable indicating membership in the road or off-road stratum (STRATA). Strata was the only categorical fixed effect variable. All others were continuous.

### Covariance Structures

Because quadrangle maps (QUAD) were randomly selected for subsampling from a the statewide population of quadrangles, these were included as random effects in the GLMM. Three covariance models for  $G$  were considered (table 2).

**Table 2.—Covariance structures for  $G$  and  $R$ .**

Structure	Form
Simple	$G_{ij} = \sigma^2$ for $i = j$ , else 0 <span style="float: right;">(10)</span>
Compound symmetry	$G_{ij} = \sigma_1^2 + \sigma^2$ for $(i = j)$ , else $\sigma_1^2$ <span style="float: right;">(11)</span>
Varying coefficients	$G_{ij} = \sigma_{ij}^2$ for $(i = j)$ , else 0 <span style="float: right;">(12)</span>
Spherical spatial	$R_{ij} = \sigma_e^2 \left[ 1 - (3d_{ij}/2\rho) + (d_{ij}^3/2\rho^3) \right]$ for $d_{ij} \leq \rho$ , else 0 <span style="float: right;">(13)</span>

A spherical spatial covariance structure, illustrated in the last row of table 2, was considered for  $R$ . Here covariance between sample points is modeled as a function of distance between those points, accounting for both correlation between clustered locations and potential correlation between sample points in relatively close quadrangles. Although numerous spatial structures could have been tried, the spherical structure is very flexible and converged more readily in preliminary trials.

## Model Fitting Strategy

Parameters in the GLMM were estimated through pseudo-likelihood procedures as described in Wolfinger and O'Connell (1993) using a SAS macro supplied by Russ Wolfinger of SAS Institute Inc. This macro uses PROC MIXED and the Output Delivery System, requiring SAS/STAT and SAS/IML release 6.08 or later.

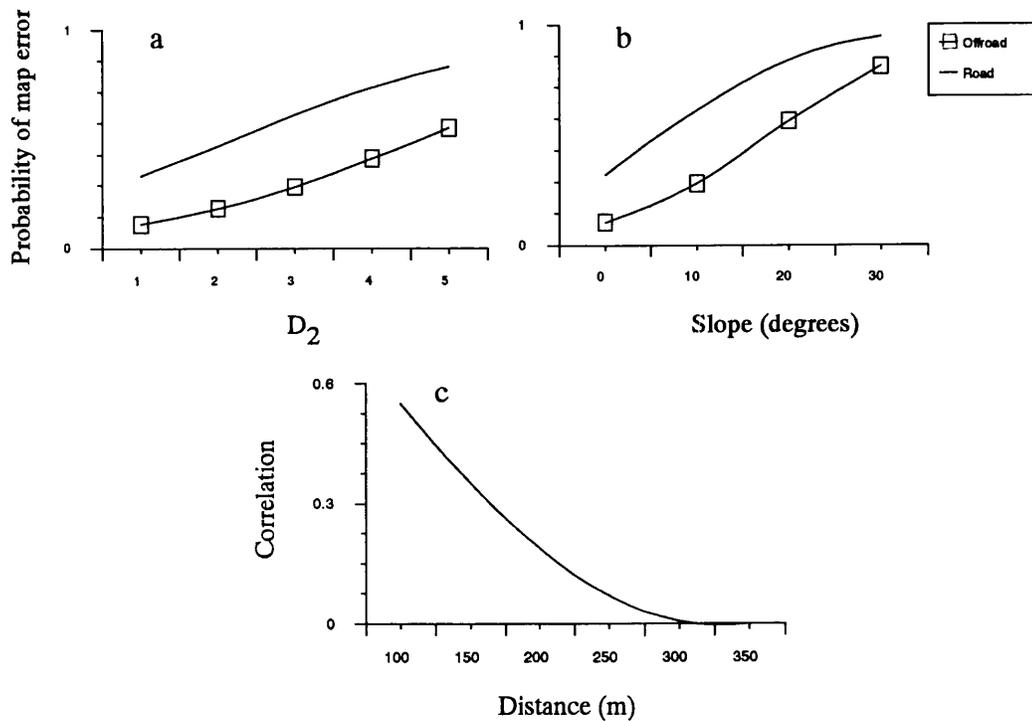
An iterative model fitting strategy was adopted. After identifying quadrangle maps as our random effects, all fixed effects were included in the model. We tried all combinations of covariance structures for  $\mathbf{G}$  as listed in table 2. The best covariance structure was selected based on Akaike's Information Criterion and Schwarz's Bayesian Criterion (Wolfinger 1993). In subsequent iterations fixed effects were dropped based on significance of parameter estimates, likelihood ratio tests, and predictive capability of the model. Again, best covariance structure was selected for each iteration. Because of collinearity, the four measures of heterogeneity were considered in the model separately.

## RESULTS

Likelihood ratio tests and parameter significance levels led us to favor a parsimonious model containing only STRATA, SLOPE and  $D_2$  as fixed effects. Exclusion of other variables had little impact on the predictive capability of the model based on confusion matrices and plots of predicted values from different model trials. The covariance structures selected were simple and spherical for  $\mathbf{G}$  and  $\mathbf{R}$ , respectively. The signs of the fixed effects parameters indicate the relationship between error and the variables (table 3). In this case, positive values for SLOPE and  $D_2$  indicate that error increases as SLOPE and  $D_2$  increase (figures 1a -b). In contrast, the negative value for the off-road stratum indicates that probability of error is less in the off-road stratum and greater in the road stratum. The estimate of  $\rho$ , the spatial covariance parameter in  $\mathbf{R}$ , suggests that spatial dependence is negligible between sample units greater than 322 meters apart (figure 1c).

**Table 3.—Parameter estimates and their standard errors for final GLMM.**

Parameter	Estimate	SE	Pr > $\chi^2$ or ( $Z^*$ )
Intercept	- 1.78	0.85	0.04
STRATA (off-road)	- 1.39	0.59	0.02
SLOPE	0.12	0.07	0.08
$D_2$	0.58	0.31	0.06
$\rho$	322.46	124.86	0.01*
$\sigma^2$	1.63	1.37	0.23*



**Figure 1a-b.—Relationship between probability of map error and fixed effects. 1c.—Correlation modeled as a function of distance between sample points.**

## DISCUSSION

Frequently in broad-scale sample surveys clustered and subsampling sample designs are adopted for the sake of efficiency in estimation of population means and totals. However, such cost-effective designs can hamper efforts to further explore ecological relationships under a classical linear model framework by violating model assumptions like independence and normality of errors. In this study we illustrated how a GLMM affords the flexibility to analyze sample survey accuracy data. Through a GLMM we were able to include both fixed and random effects, account for spatially autocorrelated errors, and allow for a variety of covariance structures for the random effects.

The model we fit for blackbrush contained some information for map users beyond simple probability of misclassification by cover-type, and also provided information to the map-maker for model improvements. We learned that incorrect mapping of this cover-type tended to occur near roads in steep and heterogeneous areas. The fact that strata was a significant contributor to our model of map error could be an indication that vegetation away from roads differs from that near roads, and is governed by an environmental factors not accounted for in the other

fixed effects. However, the better performance in the off-road strata could be an artifact of different quality in data collection efforts between the two strata. For example, georeferencing data in linear clusters may have been easier or done with greater care, making off-road data less subject to positional error.

The inclusion of slope in our model might highlight the difficulty of classification in steep and often shadowy areas. Slope was not included in the initial classification model for blackbrush in the Colorado Plateau, and its inclusion might improve maps of that cover-type. The notion that classification in heterogeneous areas is tougher than in homogenous areas is not new, but our model helps determine the magnitude of heterogeneity's contribution to error. Also, numerous indices of heterogeneity are available in the literature and we illustrated that some indices may make a more significant contribution than others to models of map error. Map users might also find the model results helpful, making them more skeptical of mapped blackbrush on steeper slopes and in more heterogeneous areas.

### ACKNOWLEDGMENTS

We would like to thank our reviewers for their thoughtful comments. We are also grateful to Ron Tymcio who tackled many GIS challenges, David Early for data collection, Collin Homer who served as our patient and ever-present practical conscience, and especially Scott Bassett, who can program anything, fast.

### REFERENCES

- Edwards, T. C., Jr.; Moisen, G. G.; Cutler, D. R. 1996. Assessing map accuracy in an ecoregion-scale cover map. In review, *Ecological Applications*.
- Hill, M. O. 1973. Diversity and evenness: a unifying notation and its consequences. *Ecology*. 54: 427-432.
- Homer, C.; Ramsey, R. D.; Edwards, T. C., Jr.; Falconer, A. In press. Landscape cover-type modelling using a multi-scene TM mosaic. *Photogrammetric Engineering and Remote Sensing*.
- Ludwig, J. A.; Reynolds J. F. 1988. *Statistical ecology*. New York: John Wiley and Sons. 337 p.
- Moisen, G. G.; Edwards, T. C. Jr.; Cutler, D. R. 1994. Spatial sampling to assess classification accuracy of remotely sensed data. In: Michener, W. K.; Brunt, J. W.; Stafford, S. G., eds. *Environmental Information Management and Analysis: Ecosystem to Global Scales*. London: Taylor and Francis: 159-176.
- Omernik, J. M. 1987. Map supplement: ecoregions of the conterminous United States. *Annals of the Association of American Geographers*. 77: 118-125 (map).
- Roberts, D. W.; Cooper, S. V. 1989. Concepts and techniques of vegetation mapping. In: Ferguson, D.; Morgan, P.; Johnson, F. D., eds. *Land Classifications based on vegetation: applications for resource management*.

- Gen. Tech. Rep. INT-257. Ogden, UT: U. S. Department of Agriculture, Forest Service, Intermountain Research Station: 90-96.
- Simpson. E. H. 1949. Measurement of diversity. *Nature*. 163: 688.
- Wolfinger, R. D. 1993. Covariance structure selection in general mixed models. *Communications in Statistics: Simulation and Computation* 22(4): 1079-1106.
- Wolfinger, R. D.; O'Connell, M. 1993. Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* . 48: 233-243.

### **BIOGRAPHICAL SKETCH**

Gretchen Moisen is a Research Forester with the Interior West Resource Inventory and Monitoring Unit of the Forest Service's Intermountain Research Station. She is currently working toward her Ph.D. in Statistics at Utah State University. In addition to her work on map accuracy, Gretchen is involved in sample design and broad-scale vegetation modeling studies.

Richard Cutler is an Associate Professor in the Department of Mathematics and Statistics at Utah State University. He received his Ph.D. in Statistics from University of California, Berkeley in 1989 and his research interests include sample survey theory, census adjustment, generalized linear models and robust estimation.

Tom Edwards is a Research Ecologist with the National Biological Service's Utah Cooperative Fish and Wildlife Research Unit, and is an Assistant Professor in the Department of Fisheries and Wildlife, Utah State University. Tom received his Ph.D. in wildlife biology from the University of Florida in 1987. His research interests include wildlife habitat relations modeling and the development of methods for assessing and monitoring biological diversity at large landscape scales.

# Statistical Properties of Measures of Association and the Kappa Statistic for Assessing the Accuracy of Remotely Sensed Data Using Double Sampling

Mohammed A. Kalkhan<sup>1</sup>, Robin M. Reich<sup>2</sup>, and  
Raymond L. Czaplewski<sup>3</sup>

---

**Abstract.**--A Monte Carlo simulation was used to evaluate the statistical properties of measures of association and the Kappa statistic under double sampling with replacement. Three error matrices representing three levels of classification accuracy of Landsat TM Data consisting of four forest cover types in North Carolina. The overall accuracy of the five indices ranged from 0.35% to 82.1% depending on the number of classes, the level of classification accuracy of satellite imagery, and the simulated sample sizes of reference plots. Statistical criteria used in the evaluation included: percent bias, mean squared error, relative error, ratio of the mean variance to the simulation variance, and 95% confidence coverage rates. Results of the simulation indicated that double sampling provided unbiased estimates of the overall accuracy of remotely sensed imagery irrespective of the number of classes in the image being analyzed, or sample size. While no one index was superior for all levels of accuracy, number of classes, or sample size, the Kappa statistic and Pearson's P provided the best estimates of the overall accuracy of the remotely sensed images. Results from previous studies suggest that increasing the sample size, or reducing the number of classification in the remotely sensed image may increase the accuracy and precision of the estimates. However, this was not the case in this study because the low accuracy of the aerial photos to ground data.

---

---

<sup>1</sup> *Biometrician, Natural Resource Ecology Laboratory, Colorado State University, Fort Collins, CO 80523, USA.*

<sup>2</sup> *Associate Professor, Department of Forest Sciences, Colorado State University, Fort Collins, CO 80523, USA.*

<sup>3</sup> *USDA Forest Service Research, Rocky Mountain Forest and Experiment Station, Fort Collins, CO 80526, USA.*

## INTRODUCTION

High resolution satellite data (*HRSD*), such as Landsat Thematic Mapper (TM) Data, the Systeme Probatoire D'Observation de la Terre (SPOT-1,2: France) satellite imagery, and JERS-1,OPS (Japan) provide natural resource managers with vital information for mapping, monitoring, and assessing the condition of their resources (Kalkhan 1994). Manager are also concerned with assessing the accuracy of their thematic maps. The high cost of collecting ground data limits the sample sizes used in assessing the accuracy of a thematic map. To reduce this cost, one could use aerial photographs. Aerial photographs are less expensive than ground data, but their degree of misclassification tends to be higher (Kalkhan *et al.* 1995b). As an alternative, some combination of aerial photography and ground data (i.e., double sampling) could be used to improve the accuracy assessment of the remotely sensed data at a lower cost.

### Measures of Association and the Kappa Statistic

Indices developed by Turk (1979), Hellden (1980), and Short (1982) estimate the accuracy of individual categories of remotely sensed data based on the errors of commission and omission associated with the estimated error matrix. A major disadvantage of these indices is that they tend to underestimate the accuracy (Rosenfield and Fitzpatrick-Lins 1986). Another disadvantage is that no valid variance formula has been developed for these indices, making it difficult to evaluate the precision of the estimates (Kalkhan 1994). Because of this, these indices are not often used to assess the accuracy of the remotely sensed images. In addition, users of remotely sensed images are more interested in evaluating the overall accuracy of their thematic map not just individual categories (Kalkhan 1994). As an alternative, users of remotely sensed images may prefer to use indices such as Pearson's coefficient of mean square contingency ( $Q^2$ ), Pearson's P, Tschuprow's T, or Cramer's V, for assessing accuracy. These latter indices are related to the Chi-square statistic and have more desirable statistical properties than the other indices (Kalkhan 1994). In addition, Bishop *et al.* (1975, p 385-386), Kalkhan (1994), and Kalkhan *et al.* (1995a) provide detailed information about the measures of association and evaluations of their statistical properties.

The Kappa statistic ( $K$ ) was originally developed by Cohen (1960) to measure the observed agreement for categorical data (Landis and KocK 1977). Since then, the Kappa statistic has received considerable attention in remote sensing applications (Congalton and Mead 1983, Rosenfield and Fitzpatrick-Lins 1986, Hudson and Ram 1987, Stehman 1992, Czaplewski 1994, Kalkhan 1994). The Kappa statistic was described in detail by Bishop *et al.* (1975, p. 395-400) and Kalkhan (1994).

## **Double Sampling and Accuracy Assessments**

Kalkhan *et al.* (1995b) point out that double sampling and accuracy assessments involve three phases: (1) unbiased sampling of the landscape; (2) accuracy assessments using a composite estimator; and (3) the integration of data and accuracy assessments into a geographic information systems (GIS). In this paper we discuss phase 2.

Estimating the accuracy of maps derived from remotely sensed data requires statistical sampling of aerial photographs or ground reference plots to insure reliable estimates (Card 1982). However, collecting information to assess the accuracy of a thematic map with respect to reference data is time consuming and costly. In designing a cost-effective sampling method, two feats must be accomplished, both of which affect the quality of information contained in the sample and the precision of the estimate. First, an unbiased sample must be selected from the population. Second, the amount of variation in the sample statistic must be controlled somewhat by the sampling method (Scheaffer *et al.* 1990). Since field observations are expensive, a design that provides a precise estimate of a parameter at minimum cost is desirable (Cochran 1977).

Two phases are required in double sampling. The first phase is to locate a sample point on an aerial photograph. It is preferable to pinpoint a location in the center of the aerial photo because it has less error and distortion (Kalkhan *et al.* 1995b). Next, the sampling point is georeferenced to the Landsat TM imagery through x,y coordinates. In the second phase, the point selected from the aerial photo is located on the ground to verify cover type mapping accuracy.

### **Accuracy Assessment: The Composite Estimator**

Kalkhan *et al.* (1995b) point out that validating the accuracy of coarse- and fine-scale vegetation maps using double sampling requires accurate estimates of bias and variance at multiple spatial scales. For this, a composite estimator developed by Mayback (1979) can be used. The composite estimator is based on data from two phases of sampling, i.e. remotely sensed data and field data. Czaplowski (1992) proposed to use the composite estimator as a statistical method to improve the precision of the estimates. The composite estimator provides considerable flexibility to accommodate complex sampling designs for cross-classified census and sample survey data (Czaplowski 1995). Adding aerial photography as an intermediate phase creates the "double sampling." More details about the composite estimator and its uses with double sampling were provided by Kalkhan *et al.* (1995b).

In assessing the classification accuracy of Landsat TM Data, we used an error matrix "contingency table" which is the most common way to represent the classification accuracy of remotely sensed data or thematic maps. The error matrix consists of a series of rows and columns containing the number of sample units, such as pixels. Each unit is assigned to a particular category relative to its actual type based on a set of reference data. The error matrix provides the user with information on the accuracy of individual categories, and both errors of commission and omission in the classification (Rosenfield and

Fitzpatrick-Lins 1986). Errors of commission relate to user accuracies, while errors of omission represent the accuracy of the remotely sensed data.

The objective of this study is to evaluate the statistical properties of Pearson's coefficient of mean square contingency ( $Q^2$ ), Pearson's P, Tschuprow's T, Cramer's V, and the Kappa statistic for assessing the classification accuracy of remotely sensed data using double sampling.

## METHODS

The data used in this study were from a pilot study designed to evaluate the use of Landsat TM data in identifying forest cover types in the state of North Carolina (Kalkhan 1994). A subsample of data was used to construct three composite error matrices representing three levels of classification accuracy (poor, average, and good) with four forest cover types. Forest cover types were identified using an unsupervised classification procedure available in *ERDAS Software* using bands TM3 (red), TM4 (IR1), and TM5 (IR2). Two error matrices were required for assessing the accuracy of thematic maps using the Kappa statistic and measures of association with double sampling. The first error matrix contained information from Landsat TM data and aerial photos, while the second error matrix contained information from aerial photos and ground data. Each error matrix contained count data (i.e. pixel, spatial information) that were converted to a joint probability matrix for the purpose of sampling and evaluating the accuracy of the maps (Kalkhan 1994). Each error matrix was estimated using Monte Carlo simulation (Rubinstein 1981) using double sampling with replacement. The error matrix relating the aerial photo to ground data was estimated using sample sizes of 50, 100, 150, 300, and 800. The error matrix relating the remotely sensed imagery to the aerial photo was estimated using sample sizes ranged from 200 to 25600 depending on number ground data used in the second phase of sampling. This process was repeated 5000 times using a Monte Carlo simulation based on a *GAUSS* matrix language computer program. The bias for the five indices was computed as the difference between the average estimates of the measures of association and the Kappa statistic and their true population mean which is known without error in the Monte Carlo simulation.

In addition to knowing whether an estimate is unbiased, it is desirable to know something about the precision of the estimates. One method of comparing the precision of estimates with different amounts of bias is to use the mean squared error (MSE). Cochran (1977) defines mean square error as the "Variance + Bias<sup>2</sup>". Therefore, an estimator with a smaller MSE is considered more precise than one with a larger MSE, even though the latter may have a smaller variance (Reich and Hussin 1993).

Because of the possible effect of bias on the observed coverage rates, one can estimate the impact of this bias by examining the relative error. The relative error is defined as the ratio of bias to the square root of the mean square error. A large relative error has the effect of distorting confidence probabilities (Cochran 1977). For example, the actual confidence level associated with a nominal 95% confidence interval for an estimate with a

relative error of 0.20 is approximately 0.9454 (Reich and Hussin 1993). Absolute relative errors less than 0.15 have little impact on coverage rates.

In addition to a bias in estimating the overall accuracy, there may also be a bias associated with estimating the variance. This can affect the observed coverage rates. One way of evaluating this bias is to compute the ratio of the mean variance to the simulation variance. The mean variance was computed by averaging the 5000 estimated variances using the classical variance formulae found in the literature (see Kalkhan 1994). For estimating the variance of the Kappa statistic, we used Czaplewski's derived variance approximation formula (Czaplewski 1994). The simulation variance was computed as the variance of the 5000 simulated estimates of overall accuracy. A 95% confidence interval for each index was computed using the appropriate standard error of the estimate, and the proportion of confidence intervals enclosing the true population mean was determined.

## RESULTS AND DISCUSSION

The observed (true) values of overall accuracy for the remotely sensed images ranged from 0.33 to 0.87, while the overall accuracy of aerial photo to ground was 0.79 (Table 1). The observed overall accuracy for the Kappa statistic and measures of association ranged from 0.004 to 2.03.

The scenes with four classes had a mean accuracy for the Kappa statistic ranging from -0.8 to 64.3, while Pearson's  $Q^2$ , had an overall accuracy ranging from 0.9 to 140.1 (Table 2). Pearson's P ranged from 9.4 to 75.7, while Tschuprow's T and Cramer's V had an overall accuracy ranging from 5.5 to 68.

Note that the overall accuracy of the Kappa statistic increased for poor classification as the aerial photo to ground sample sizes increased, while for the measures of association their overall accuracy decreased as the sample sizes of remotely sensed images increased. For average and good classifications the overall accuracy decreased for all indices when the aerial photo to ground sample sizes increased. One could expect that the overall accuracy of the aerial photo to ground would be higher compared to the accuracy of remotely sensed images. The Landsat TM imagery was taken on October 8, 1985, while the aerial photos were taken on October 10 and November 9 of 1985. The ground survey to collect ground sampling points was accomplished during November 11, 1982 through November 1986. This might have an effect on assessing the accuracy. Increasing the sample size did not improve the overall accuracy. In fact, the overall accuracy decreased as sample size increased.

### Relative bias

A consistent trend was observed in the relative bias with respect to the accuracy of the scene and number of samples at different phases of sampling (Table 3). Pearson's P provided the best estimate of overall accuracy ranging from -0.1 to -0.8 (Table 3) followed by Tschuprow's T, Cramer's V, and Pearson's  $Q^2$  with relative bias ranging from -0.2 to -1.0.

Table 1.--Observed values of five accuracy assessment indices.

Index <sup>a</sup>	Poor	Average	Good	Photo-Ground
O	0.3221	0.5594	0.8744	0.7884
K	0.0035	0.3571	0.8206	0.6967
Q	0.2308	0.9447	2.0320	1.4862
P	0.4330	0.6970	0.8186	0.7732
T	0.2774	0.5613	0.8230	0.7039
V	0.2774	0.5613	0.8230	0.7039

<sup>a</sup> Overall accuracy (O), Kappa statistic(K), Pearson's coefficient of mean contingency (Q<sup>2</sup>), Pearson's (P), Tschaprowski's (T), and Cramer's (V).

Table 2.--Estimated<sup>1</sup> values of five accuracy assessment indices.

Index <sup>a</sup> M	Poor	Average	Good
K 50	-0.7 -0.8 -0.7 -0.8A	30.3 30.3 30.0 29.6	64.1 64.0 64.3 64.2
Q	14.5 13.6 13.3 13.2	44.8 44.8 43.2 42.4	139.0 138.8 140.1 139.9
P	34.3 33.4 33.1 32.9	54.7 54.7 54.1 53.8	75.6 75.5 75.7 75.7
T	21.4 20.8 20.5 20.4	38.4 38.4 37.7 37.4	67.7 67.6 67.9 67.9
V	21.6 20.9 20.7 20.6	38.6 38.6 37.9 37.7	67.8 67.7 68.0 68.0
K 100	-0.4 -0.4 -0.6 -0.5B	28.4 28.1 28.1 28.2	62.2 62.3 62.0 62.3
Q	7.0 6.7 6.7 6.5	34.5 33.9 33.6 33.7	125.4 125.8 125.4 126.1
P	24.8 24.3 24.2 23.9	50.1 49.8 49.7 49.7	74.2 74.3 74.2 74.3
T	14.9 14.6 14.5 14.3	33.7 33.4 33.3 33.3	64.4 64.5 64.4 64.6
V	14.9 14.6 14.5 14.3	33.7 33.4 33.3 33.3	64.4 64.5 64.4 64.6
K 150	-0.3 -0.4 -0.3 -0.4C	27.9 27.7 27.6 27.6	61.5 61.3 61.3 61.4
Q	4.6 4.4 4.4 4.3	31.0 30.5 30.3 30.2	120.4 119.6 119.7 120.2
P	20.4 20.0 19.9 19.7	48.3 48.0 47.7 47.4	73.6 73.5 73.6 73.6
T	12.1 11.8 11.7 11.6	32.0 31.7 31.6 31.5	63.2 63.0 63.0 63.1
V	12.1 11.8 11.7 11.6	32.0 31.7 31.6 31.5	63.2 63.0 63.0 63.1
K 300	-0.2 -0.2 -0.2 -0.2D	27.5 27.3 27.3 27.3	60.3 60.3 60.3 60.1
Q	2.3 2.2 2.2 2.1	27.9 27.5 27.4 27.3	113.1 112.9 113.0 112.3
P	14.6 14.3 14.1 14.1	46.5 46.3 46.2 46.2	72.7 72.7 72.7 72.6
T	8.6 8.4 8.2 8.2	30.4 30.2 30.2 30.1	61.3 61.2 61.3 61.1
V	8.6 8.4 8.2 8.2	30.4 30.2 30.2 30.1	61.3 61.2 61.3 61.1
K 800	-0.1 -0.1 -0.1 -0.0E	27.3 27.3 27.2 27.3	59.6 59.5 59.5 59.4
Q	1.0 1.0 1.0 0.9	26.2 26.0 25.9 26.0	108.6 108.1 108.2 107.8
P	9.9 9.7 9.5 9.4	45.5 45.4 45.3 45.3	72.1 72.0 72.0 72.0
T	5.7 5.6 5.5 5.5	29.5 29.4 29.4 29.4	60.1 60.0 60.0 59.9
V	5.7 5.6 5.5 5.5	29.5 29.4 29.4 29.4	60.1 60.0 60.0 59.9

<sup>1</sup> Estimated values were multiplied by 100.

<sup>a</sup> See Table 1 for description of variables.

N Sample size from aerial photo to ground (Second phase of sampling).

A Sample sizes from remotely sensed to aerial photo are 200,400,800,1600.

B Sample sizes from remotely sensed to aerial photo are 400,800,1600,3200.

C Sample sizes from remotely sensed to aerial photo are 600,1200,2400,4800.

D Sample sizes from remotely sensed to aerial photo are 1200,2400,4800,9600.

E Sample sizes from remotely sensed to aerial photo are 3200,6400,12800,25600.

F First phase of sampling from remotely sensed image to aerial photo.

Table 3.--Percent bias\* for five accuracy indices.

Index <sup>a</sup> M	Poor	Average	Good
K 50	-3.1 -3.4 -3.0 -3.3A	-0.2 -0.2 -0.2 -0.2	-0.2 -0.2 -0.2 -0.2
Q	-0.4 -0.4 -0.4 -0.4	-0.5 -0.5 -0.5 -0.6	-0.3 -0.3 -0.3 -0.3
P	-0.2 -0.2 -0.2 -0.2	-0.3 -0.3 -0.3 -0.3	-0.1 -0.1 -0.1 -0.1
T	-0.2 -0.3 -0.3 -0.3	-0.3 -0.3 -0.3 -0.3	-0.2 -0.2 -0.2 -0.2
V	-0.2 -0.2 -0.2 -0.3	-0.3 -0.3 -0.3 -0.3	-0.2 -0.2 -0.2 -0.2
K 100	-2.1 -2.2 -2.8 -2.4B	-0.2 -0.2 -0.2 -0.2	-0.2 -0.2 -0.2 -0.2
Q	-0.7 -0.7 -0.7 -0.7	-0.6 -0.6 -0.6 -0.6	-0.4 -0.4 -0.4 -0.4
P	-0.6 -0.6 -0.6 -0.6	-0.3 -0.3 -0.3 -0.3	-0.1 -0.1 -0.1 -0.1
T	-0.5 -0.5 -0.5 -0.5	-0.4 -0.4 -0.4 -0.4	-0.2 -0.2 -0.2 -0.2
V	-0.5 -0.5 -0.5 -0.5	-0.4 -0.4 -0.4 -0.4	-0.2 -0.2 -0.2 -0.2
K 150	-1.8 -2.0 -2.0 -2.2C	-0.2 -0.2 -0.2 -0.2	-0.3 -0.3 -0.3 -0.3
Q	-0.8 -0.8 -0.8 -0.8	-0.7 -0.7 -0.7 -0.7	-0.4 -0.4 -0.4 -0.4
P	-0.5 -0.5 -0.5 -0.5	-0.3 -0.3 -0.3 -0.3	-0.1 -0.1 -0.1 -0.1
T	-0.4 -0.4 -0.4 -0.4	-0.3 -0.3 -0.3 -0.3	-0.1 -0.1 -0.1 -0.1
V	-0.6 -0.6 -0.6 -0.6	-0.4 -0.4 -0.4 -0.4	-0.2 -0.2 -0.2 -0.2
K 300	-1.5 -1.6 -1.6 -1.5D	-0.2 -0.2 -0.2 -0.2	-0.3 -0.3 -0.3 -0.3
Q	-0.9 -0.9 -0.9 -0.9	-0.7 -0.7 -0.7 -0.7	-0.4 -0.4 -0.4 -0.4
P	-0.7 -0.7 -0.7 -0.7	-0.3 -0.3 -0.3 -0.3	-0.1 -0.1 -0.1 -0.1
T	-0.7 -0.7 -0.7 -0.7	-0.3 -0.3 -0.3 -0.3	-0.1 -0.1 -0.1 -0.1
V	-0.7 -0.7 -0.7 -0.7	-0.5 -0.5 -0.5 -0.5	-0.3 -0.3 -0.3 -0.3
K 800	-1.2 -1.3 -1.1 -1.1E	-0.2 -0.2 -0.2 -0.2	-0.3 -0.3 -0.3 -0.3
Q	-1.0 -1.0 -1.0 -1.0	-0.7 -0.7 -0.7 -0.7	-0.5 -0.5 -0.5 -0.5
P	-0.8 -0.8 -0.8 -0.8	-0.3 -0.3 -0.3 -0.3	-0.1 -0.1 -0.1 -0.1
T	-0.8 -0.8 -0.8 -0.8	-0.3 -0.3 -0.3 -0.3	-0.1 -0.1 -0.1 -0.1
V	-0.8 -0.8 -0.8 -0.8	-0.3 -0.3 -0.3 -0.3	-0.1 -0.1 -0.1 -0.1

\* Percent bias divided by 100.

<sup>a</sup> See Table 1 and 2 for description of variables and sample sizes.

Table 4.--Mean square error\* of five accuracy indices.

Index <sup>a</sup> M	Poor	Average	Good
K 50	954.3 1179.8 894.8 1077.7A	2.3 2.3 2.6 2.9	4.8 4.8 4.7 4.7
Q	13.8 16.7 17.9	18.2 27.7 27.7 29.5	30.3 10.0 10.1 9.6
P	4.3 5.3 5.6	5.8 4.6 4.6 5.0	5.2 0.6 0.6 0.6
T	5.2 6.3 6.7	6.9 10.0 10.0 10.8	11.1 3.2 3.2 3.1
V	5.0 6.1 6.5	6.7 9.7 9.7 10.5	10.8 3.1 3.1 3.0
K 100	441.6 500.0 789.1 561.3B	4.1 4.5 4.5 4.5	5.9 5.8 6.0 5.8
Q	48.3 50.1 50.2	51.7 40.3 41.1 41.5	41.4 14.7 14.5 14.7
P	18.2 19.2 19.4	20.1 7.9 8.1 8.2	8.2 0.9 0.9 0.9
T	21.4 22.6 22.7	23.6 16.0 16.0 16.4	16.6 16.5 4.7 4.7
V	21.4 22.6 22.7	23.6 16.0 16.0 16.4	16.6 16.5 4.7 4.7
K 150	319.2 412.3 392.3 468.9C	4.8 5.1 5.2 5.2	6.3 6.4 6.4 6.4
Q	64.0 65.3 65.5	66.2 45.1 45.9 46.1	46.4 16.6 16.9 16.9
P	28.0 29.1 29.3	29.8 9.4 9.7 9.8	9.9 1.0 1.0 1.0
T	31.9 33.0 33.2	33.8 18.5 18.9 19.0	19.2 5.4 5.5 5.5
V	31.9 33.0 33.2	33.8 18.5 18.9 19.0	19.2 5.4 5.5 5.5
K 300	225.3 272.0 257.1 238.8D	5.3 5.5 5.5 5.5	7.0 7.1 7.1 7.2
Q	80.9 81.6 82.2	82.3 49.6 50.3 50.4	50.5 19.6 19.8 19.7
P	43.8 44.7 45.4	45.6 11.0 11.3 11.3	11.4 1.2 1.3 1.3
T	47.8 48.7 49.4	49.5 21.0 21.3 21.4	21.5 6.5 6.5 6.5
V	47.8 48.7 49.4	49.5 21.0 21.3 21.4	21.5 6.5 6.5 6.5
K 800	144.1 167.7 131.4 130.3E	5.5 5.6 5.7 5.6	7.5 7.5 7.5 7.6
Q	91.3 91.5 91.8	92.0 52.3 52.5 52.7	52.6 21.7 21.9 21.9
P	59.6 60.2 60.9	61.1 12.1 12.2 12.2	12.2 1.4 1.4 1.5
T	63.0 63.5 64.1	64.4 22.5 22.6 22.7	22.7 7.3 7.3 7.3
V	63.0 63.5 64.1	64.4 22.5 22.6 22.7	22.7 7.3 7.3 7.3

\* MSE divided by 100.

<sup>a</sup> See Table 1 and 2 for description of variables and sample sizes.

Table 5.--Nominal\* 95% confidence coverage rates using the classical variance formula for five accuracy indices.

Index <sup>a</sup> M	N	Poor					Average					Good				
		O	K	P	T	V	O	K	P	T	V	O	K	P	T	V
50	200	41	80	49	15	73	3	66	4	1	10	15	12	19	4	32
	400	26	66	29	8	45	1	50	2	0	4	9	6	11	18	
	800	17	50	18	6	29	1	36	1	0	2	7	4	7	21	
	1600	11	36	11	3	18	0	25	0	0	1	4	3	4	2	7
100	400	2	81	3	0	9	0	45	0	0	0	1	2	0	4	
	800	1	67	1	0	3	0	28	0	0	0	1	0	0	2	
	1600	1	51	1	0	2	0	18	0	0	0	0	0	0	0	
	3200	0	39	0	0	1	0	13	0	0	0	0	0	0	0	
150	600	0	82	0	0	1	0	30	0	0	0	0	0	0	0	
	1200	0	67	0	0	0	0	16	0	0	0	0	0	0	0	
	2400	0	52	0	0	0	0	5	0	0	0	0	0	0	0	
	4800	0	39	0	0	0	0	0	0	0	0	0	0	0	0	
300	1200	0	67	0	0	0	0	1	0	0	0	0	0	0	0	
	2400	0	54	0	0	0	0	1	0	0	0	0	0	0	0	
	4800	0	38	0	0	0	0	1	0	0	0	0	0	0	0	
	9600	0	7	0	0	0	0	0	0	0	0	0	0	0	0	
800	3200	0	67	0	0	0	0	0	0	0	0	0	0	0	0	
	6400	0	53	0	0	0	0	0	0	0	0	0	0	0	0	
	12800	0	40	0	0	0	0	0	0	0	0	0	0	0	0	
	25600	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

\* 95% Confidence Interval multiply by 100.

<sup>a</sup> See Table 1 for description of variables.

<sup>b</sup> Sampling from remote sensing image to aerial photo.

Table 6.--Relative error\* for five accuracy assessment indices.

Index <sup>a</sup> M	Poor	Average	Good
K 50	-0.0 -0.0 -0.0 -0.0A	-3.6 -3.6 -3.6 -3.6	-8.2 -8.2 -8.2 -8.2
Q	-2.3 -2.3 -2.3 -2.3	-9.4 -9.4 -9.4 -9.4	-20.3 -20.3 -20.3 -20.3
P	-4.3 -4.3 -4.3 -4.3	-7.0 -7.0 -7.0 -7.0	-8.2 -8.2 -8.2 -8.2
T	-2.8 -2.8 -2.8 -2.8	-5.6 -5.6 -5.6 -5.6	-8.2 -8.2 -8.2 -8.2
V	-2.8 -2.8 -2.8 -2.8	-5.6 -5.6 -5.6 -5.6	-8.2 -8.2 -8.2 -8.2
K 100	-0.0 -0.0 -0.0 -0.0B	-3.6 -3.6 -3.6 -3.6	-8.2 -8.2 -8.2 -8.2
Q	-2.3 -2.3 -2.3 -2.3	-9.4 -9.4 -9.4 -9.4	-20.3 -20.3 -20.3 -20.3
P	-4.3 -4.3 -4.3 -4.3	-7.0 -7.0 -7.0 -7.0	-8.2 -8.2 -8.2 -8.2
T	-2.8 -2.8 -2.8 -2.8	-5.6 -5.6 -5.6 -5.6	-8.2 -8.2 -8.2 -8.2
V	-2.8 -2.8 -2.8 -2.8	-5.6 -5.6 -5.6 -5.6	-8.2 -8.2 -8.2 -8.2
K 150	-0.0 -0.0 -0.0 -0.0C	-3.6 -3.6 -3.6 -3.6	-8.2 -8.2 -8.2 -8.2
Q	-2.3 -2.3 -2.3 -2.3	-9.4 -9.4 -9.4 -9.4	-20.3 -20.3 -20.3 -20.3
P	-4.3 -4.3 -4.3 -4.3	-7.0 -7.0 -7.0 -7.0	-8.2 -8.2 -8.2 -8.2
T	-2.8 -2.8 -2.8 -2.8	-5.6 -5.6 -5.6 -5.6	-8.2 -8.2 -8.2 -8.2
V	-2.8 -2.8 -2.8 -2.8	-5.6 -5.6 -5.6 -5.6	-8.2 -8.2 -8.2 -8.2
K 300	-0.0 -0.0 -0.0 -0.0D	-3.6 -3.6 -3.6 -3.6	-8.2 -8.2 -8.2 -8.2
Q	-2.3 -2.3 -2.3 -2.3	-9.4 -9.4 -9.4 -9.4	-20.3 -20.3 -20.3 -20.3
P	-4.3 -4.3 -4.3 -4.3	-7.0 -7.0 -7.0 -7.0	-8.2 -8.2 -8.2 -8.2
T	-2.8 -2.8 -2.8 -2.8	-5.6 -5.6 -5.6 -5.6	-8.2 -8

The Kappa statistic had high relative bias ranging from -0.2 to -3.4. As the overall accuracy of the scene increased, the relative bias decreased for all indices. The relative bias for the Kappa statistic decreased when the aerial photo to ground sample sizes increased for poor classification. The opposite was found for the measures of association for average and good classification; the bias increased as the sample size increased.

### **Mean squared error**

Pearson's P had the smallest MSE across all sample sizes ranging from 0.6 to 61.1 followed by Cramer's V and Tschuprow's T with the MSE ranging from 3.0 to 64.4 for all levels of classification (Table 4). Pearson's Q<sup>2</sup> had a MSE ranging from 9.7 to 92, while the Kappa statistic had the largest MSE. In general, the mean square error decreased with increased accuracy of the remotely sensed image. The MSE for the Kappa statistic decreased as the aerial photo to ground sample sizes increased, while the opposite was observed for the measures of association. The MSE increased consistently for all indices as the sample size associated with estimating the error matrix relating that to aerial photo to the ground. For the poor classification, the Kappa statistic decreased as the sample size of the photo-ground increased.

### **95 Percent coverage rates**

None of the five indices provided consistent coverage rates across all sample sizes and levels of accuracy (Table 5). The Kappa statistic provided reasonable coverage rates, but was not consistent for the poor classification. The coverage rate was very poor for the average and good classification. The inconsistency in coverage rates may be due to the bias, which could distort the confidence intervals. The lower than expected coverage rates may also be due to a bias in estimating the variance.

### **Bias relative to the root mean squared error**

In computing the relative errors, the simulation variance was used instead of the classical variance to facilitate comparison between all accuracy indices. Except for a few instances, all the absolute relative errors exceeded 0.20 (Table 6). This suggests that the observed biases are large enough relative to the variability to distort the expected coverage rates. This would help explain why the coverage rates were lower than expected.

### **Ratio of the mean variance to the simulation variance**

As noted, the lower than expected coverage rates also may be due to a bias associated with estimating the variance. One way to evaluate this bias is to calculate the ratio of the mean variance to the simulation variance (Table 7). It is assumed that the simulation variance provides an unbiased estimate of the variance. Thus, a ratio of less than 1 indicates an underestimation of the variance, while a ratio greater than 1 indicates an

overestimation (Cochran 1977). Using this as a guideline, one can see that all indices underestimated the variance, except for a few cases where Cramer's V overestimated the variance. There was also some indication that the bias decreased as the aerial photo to ground sample sizes increased for poor classification. In addition, as the level of the accuracy improved, the bias increased.

The biases associated with estimating the variances may be due to the asymptotic nature of the variance formulae which assume large sample sizes. When used with a small sample size the variance formulae may not provide valid estimates of the variance (see Kalkhan 1994), especially for the Kappa statistic where we used derived variance formula developed by Czaplewski (1994). Also, the variance estimates of Pearson's P, Tschuprow's T, and Cramer's V are based on a nonlinear transformation of the asymptotic variance of Pearson's  $Q^2$ . This transformation may also introduce a bias in the estimation process. In general, increasing the sample size of the remotely sensed image did not seem to decrease the bias of the mean variance estimates. This is because the overall accuracy of aerial photo to ground comparison was less than the Landsat TM imagery.

## CONCLUSION

The Kappa statistic and Pearson's P provided the best estimates of overall accuracy for all levels of classification, sample sizes, and numbers of forest cover types. In addition, the results of the Kappa statistic in this study are applicable to Pielou's index of segregation (Pielou 1960, 1977; Kalkhan *et al.* 1995a). Poor performance of the measures of association were due to large biases associated with the statistic. The ratios of the estimated mean variance to the simulation variance suggested that the classical variance formula for the various indices provided biased estimates of variance, irrespective of the number of classes, or the sample size. This bias may be due to the asymptotic nature of the variance formulae. These variance formulae may not provide valid estimates when used with small sample sizes. Some of the bias associated with the variance may be due to the nonlinear transformation used to approximate the variance of Pearson's P, Tschuprow's T, and Cramer's V. To overcome this problem, one could use bootstrapping (Efron 1979) to provide unbiased estimates of the variance (Kalkhan 1994).

Double sampling, as applied in this study, allows one to make inference about the whole image and not a preselected subset of homogenous test sites. It also suggests that, increasing the sample size, or reducing the number of classes on the remotely sensed image may increase the accuracy and precision of the estimates. However, this was not the case in this study because the accuracy of this particular aerial photo to ground was low. No matter what indices one uses to assess the accuracy of a remotely sensed image, bootstrapping should be used to ensure unbiased estimates of the variance (Kalkhan 1994). To improve the findings of this study, we suggest the use of a large number of classes for assessing the accuracy of the thematic map. Prior to assessing the accuracy of any thematic map, the user of remotely sensed must be familiar with the statistical properties of any accuracy assessment index before using it.

## REFERENCES

- Bishop, Y. M. M., S. E. Feinberg, and P. W. Hooland. 1975. Discrete multivariate analysis-theory and practice. MIT Press, Cambridge, MA. 575 p.
- Card, D. H. 1982. Using known map categorical marginal frequencies to improve map accuracy. *Photogrammetric Eng. & Remote Sensing*. 48:431-439.
- Cochran, W. G. 1977. Sampling Techniques. 3rd ed. John Wiley and Sons, New York. 428 p.
- Cohen, J. 1960. A coefficient of agreement of nominal scales. *Educ. Psychological Meas.* 20:37-46.
- Congalton, R. G., and R. A. Mead. 1983. A quantitative method to test for consistency and correctness in photointerpretation. *Photogramm. Eng. & Remote Sens.* 49:67-74.
- Czaplewski, R. L. 1992. Accuracy assessment of remotely sensed classification with multi-phase sampling and the multivariate composite estimator. P. 22 *In: 1992 (XVI<sup>th</sup>) Intern. Biometric Conference, Hamilton, New Zealand, December 7-11. Vol. 2, 268 p.*
- Czaplewski, R. L. 1994. Variance approximation for assessment of classification accuracy. Research Paper, Rocky Mountain For. & Range Exper. Station, Forest Serv. U.S.D.A., 94 p.
- Czaplewski, R. L. 1995. Assessment of classification accuracy and extent estimates for land cover map with double sampling. *Forest Science* (In Review).
- Efron, B. 1979. Bootstrap methods: Another look at the jackknife. *Annals of Stat.* 7:1-26.
- Helldon, U. 1980. A test of Landsat-2 imagery and digital data for thematic mapping illustrated by an environmental study in northern Kenya. Sweden, Lund University Natural geography Institute, Report No. 7.
- Hudson, W. D., and C. W. Ram. 1987. Correct formulation of the Kappa coefficient of agreement. *Photogramm. Eng. & Remote Sens.* 53:421-422.
- Landis, J. R., and G. G. Kock. 1977. The measurement of observer agreement for categorical data: *Biometrics*. 33:154-174.
- Kalkhan, M. A. 1994. Statistical properties of six accuracy indices using simple random and stratified random sampling: An application in remote sensing. *Ph.D. Dissertation*. Colorado State University. 134 p.
- Kalkhan, M. A., R. M. Reich, and R. L. Czaplewski. 1995a. Statistical properties of five accuracy indices in assessing the accuracy of remotely sensed data using simple random sampling. *In: 1995 ACSM/ASPRS Annual Convention & Exposition, ASPRS Technical Papers, February 27- March, 1995, Charlotte, North Carolina. P. 246-257.*
- Kalkhan, M. A., T. J. Stohlgren, and M. Coughneour. 1995b. An investigation of biodiversity and Landscape-scale gape patterns using double sampling: A GIS approach. *In: 1995 Ninth International Symposium on Geographic Information Systems for natural resources, environment and land information management, March 27-30, 1995, Vancouver, British Columbia, Canada. P. 708-712.*

- Maybeck, P.S. 1979. Stochastic Models, Estimation, and Control, vol. 1. Academic Press, New York.
- Pielou, E. C. 1961. Segregation and symmetry in two species populations as studied by nearest neighbor relations. *J. Ecology*. 49:255-269.
- Pielou, E. C. 1977. Mathematical Ecology. John Wiley and Sons, New York. 385 p.
- Reich, R. M., and Y. A. Hussin. 1993. Estimating average stand biomass for a regional forest inventory using radar backscatter. *ITC Journal*. 1:82-87.
- Rosenfield, G. H. and K. Fitzpatrick-Lins. 1986. A coefficient of agreement as a measure of thematic classification accuracy. *Photogramm. Eng. & Remote Sensing*. 52:223-227.
- Rubinstein, R. Y. 1981. Simulation and the monte carlo method. John Wiley and Sons, New York. 278 p.
- Scheaffer, R. L., W. Mendenhall, and L. Ott. 1990. Elementary Survey Sampling, 4th ed., PWS-Kent Publishing Company, Boston. MA. 390 p.
- Short, N. M. 1982. The Landsat tutorial workbook: Basis of satellite remote sensing. Greenbelt, MD., GSF Center, NASA Reference Publication 1078.
- Stehman, S. V. 1992. Comparison of systematic and random sampling for estimating the accuracy of maps generated from remotely sensed data. *Photogramm. Eng. & Remote Sens.* 58:1343-1350.
- Turk, G. 1979. GT index: A measure of the success of prediction. *Remote Sens. Environ.* 8:65-75.

### BIOGRAPHICAL SKETCH

**Mohammed A. Kalkhan** is a Research Scientist/GIS Specialist with the Natural Resource Ecology Laboratory, Colorado State University. He graduated from University of Mosul, Iraq with B.Sc. (Forestry) in 1973 and M.Sc. (Forest Mensuration) in 1980. Mohammed holds a Ph.D. in Forest Biometrics from Colorado State University (1994). Research interest include using remote sensing, GIS, biometrics, and spatial statistics to study natural resources, landscape-gap analysis and ecology, biodiversity, and accuracy assessment of remotely sensed data with sampling techniques.

**Robin M. Reich** is an Associate Professor with the Department of Forest Sciences, Colorado State University. Research interest spatial statistics, natural resource sampling and forest biometrics. Current research is aimed at integrating of spatial and geostatistical technique with GIS and remote sensing to model the spatial interaction biological populations have with themselves and their environment.

**Raymond L. Czaplewski** has spent the past 13 years working on statistical design and analysis of regional monitoring systems for forests and other environmental features, primarily the Forest Service's national Forest Inventory and Analysis Program. Other accomplishments include statistical design of the 1990 and 2000 global surveys of tropical forests by the Food and Agricultural Organization of the United Nations, and a design to monitor the effectiveness of the President's 1994 Forest Plan for the States of Oregon, Washington, and California. He has specialized in linking sample surveys to remotely sensed data and geographic information systems.

# Comparing Methods used to Combine Multi-level Samples of Compositional Data

Murray Todd Williams<sup>1</sup>

---

**Abstract.**—Accuracy assessment of categorical data with complex sampling designs is an important tool in the use of thematic maps and other applications. One important development of note is the composite estimator—an application of the Kalman filter. This paper reviews the composite estimator and compares it with standard weighted least squares regression. In many cases the two methods produce identical results, although there are situations in which one has advantages over the other. The conclusion also raises some critical questions which arise from these classical methods, and it examines contemporary analyses of compositional data. In conjunction with the methodology described here, a public-domain software application has been developed.

---

## INTRODUCTION

Much of the work discussed here emerged from natural resource issues, specifically the analysis of forest inventories carried out by the USDA Forest Service. However, the application of this work has already proven to be quite large. Let us broadly define the sort of data which can benefit from these analyses.

In single-level categorical data, measured quantities (frequently discrete) are identified as one of a group of mutually exclusive categories. One example of this is categorizing the type of forest in a section of land by tree species. Another example would be the a typical blood test in which the number of specific components (red blood cells, lymphocytes, etc.) are counted.

Multi-stage and multi-phase categorical sampling uses multiple classifiers to categorize each sampling unit. For example, in the above forest inventory situation, the composition of a section of land can be measured using three different methods: satellite imaging, aerial photography (photo interpretation), and manual observations in the field.

In this example, let a primary sampling unit be a section of land with an area of 47 acres. Table 1 shows an example where the three identification methods was used to classify each acre of land. Notice that 38 acres of land were classified

---

<sup>1</sup>*Student, Department of Statistics, Colorado State University, Ft. Collins, CO*

the same by each method. For the other 9 acres, there is some degree of disagreement between categories. (These situations are shaded in the table.) In most situations a high level of agreement is considered optimal.

Notice that Table 1 also displays the quantities of each multiple-classification as proportions of the total. This vector of proportions is referred to as a *composition*.

**Table 1.**—Mutli-level data in a typical *primary sampling unit*.

Satellite Imaging	Aerial Photography	Ground Surveys	Acres	Composition
Non-forest	Non-forest	Non-forest	12	0.255
Non-forest	Light forest	Light Forest	2	0.043
Non-forest	Light forest	Heavy forest	5	0.106
Light forest	Light forest	Light forest	15	0.319
Light forest	Heavy forest	Light forest	2	0.043
Heavy forest	Heavy forest	Heavy forest	11	0.234
<b>Total:</b>			47	1.000

There are many situations in which this type of data is encountered. Some of the possible applications become apparent:

- **Multiple classification techniques.** In these situations, various techniques might have different levels of accuracy. Such accuracy is frequently inversely proportional to cost. In the above example, satellite data might be readily available and therefore inexpensive, but the degree of accuracy obtained may be fairly poor. Ground surveys are considered to be the most accurate, but they are also exceedingly expensive.
- **Multiple time periods.** A large database of monitoring information may be available for previous time periods (say for example, a large survey taken in 1970 and 1980) where data for a recent time period (1990) is more limited. A combination of the larger two-period database may be combined with the small sample of plots which have all three time periods in order to extrapolate a better estimate of the forest composition in 1990.

### Construction of the Sample Composition Estimates

Consider an  $(n \times p)$  data matrix  $X$  with  $n$  independent primary sampling units (PSUs) and  $p$  distinct categories. Note that the category labels resemble those in Table 1. Also let each row of  $X$  represent a composition, thus summing to one. The calculations of the composition estimate  $Y$  (the sample mean of the data matrix  $X$ ) and its sample covariance are straightforward:

$$Y = \frac{1}{n} X^T \mathbf{1} \quad \text{and} \quad \Sigma_Y = \frac{1}{n} X^T (\mathbf{I} - \mathbf{1}\mathbf{1}^T) X \quad [1]$$

## COMPOSITE ESTIMATOR

The composite estimator was first applied to categorical analyses by Czaplewski (1992) as an application of the Kalman filter. Instead of motivating its derivation I will just describe its application.

Consider two composition estimates. The estimates may have any dimensionality, but for this example we will define a two-dimensional composition  $Y_1$  and a three-dimensional composition  $Y_2$ .  $Y_2$  would be a vector with the unit-sum constraint. The elements of  $Y_2$  will correspond with the rows in Table 1.

$Y_1$  will only have categories defined on only two classifications: Satellite imaging and Aerial photography. Refer to Table 2. It is necessary now to construct a binary transformation matrix which defines the unique mapping from  $Y_1$  to  $Y_2$ . Take this partition of the binary matrix and call it  $H$ . An example of  $H$  is shown as a subset of the binary matrix in Table 3.

**Table 2.**—A typical *primary sampling unit* for a two-dimensional composition  $Y_1$ .

Satellite Imaging	Aerial Photography	Acres	Composition
Non-forest	Non-forest	19	0.268
Non-forest	Light Forest	8	0.113
Light forest	Light forest	22	0.310
Light forest	Heavy forest	7	0.099
Heavy forest	Heavy forest	15	0.211
<b>Total:</b>		71	1.000

From Czaplewski (1994) the composite estimator  $Y$  can be expressed as

$$Y = Y_2 + K(Y_1 - HY_2) \quad \text{and} \quad \Sigma_Y = (I - KH)\Sigma_2 \quad [2]$$

$$\text{where} \quad K = \Sigma_2 H^T (\Sigma_1 + H \Sigma_2 H^T) \quad [3]$$

The resulting composite estimator will have the same dimension and category labels as the largest component  $Y_2$ . In a sense, the composite estimator can be considered a method of improving the estimate of  $Y_2$  by using another independent and unbiased vector estimate ( $Y_1$ ) of lower dimensionality.

A problem which can sometimes occur with both the composite estimator and the weighted least squares is that the resulting composition estimate  $Y$  can have negative components. For the composite estimator, there is a partial remedy of this problem. Multiply  $K$  in the above equations with a scalar constant  $\alpha$ .

$$K = \alpha \Sigma_2 H^T (\Sigma_1 + H \Sigma_2 H^T) \quad [4]$$

In optimal situations (where negative components do not result) we can let  $\alpha=1$ . If negative components of the composite estimator are observed, we can replace  $\alpha$  with the largest possible value in the interval  $[0,1]$  such that the resulting composite estimator is well-behaved.

## WEIGHTED LEAST SQUARES

The use of weighted (generalized) least squares allows a greater level of flexibility than the composite estimator. The only thing about the weighted least squares regression that may not be considered straightforward is the fact that the covariance matrices of the compositions will be necessarily singular. By replacing the standard matrix inverse with the generalized inverse, the problem is alleviated.

**Table 3.**—Creation of the design matrix  $X$ .

<b>WLS Regression Categories</b>			<b>Ground</b>																														
<b>Original Categories</b>			<b>Aerial</b>																														
<b>Satellite</b>	<b>Aerial</b>	<b>Ground</b>	<b>Satellite</b>																														
Non-forest	Non-forest	Non-forest	Non-forest																														
Non-forest	Light forest	Light forest	Non-forest																														
Light forest	Light forest	Light forest	Light forest																														
Light forest	Heavy forest	Light forest	Light forest																														
Heavy forest	Heavy forest	Heavy forest	Heavy forest																														
Non-forest	Non-forest	Non-forest	Non-forest																														
<missing>	Light forest	Light forest	Light forest																														
Non-forest	Light forest	Light forest	Light forest																														
Non-forest	Light forest	Heavy forest	Light forest																														
Light forest	Light forest	Light forest	Light forest																														
Light forest	Heavy forest	Light forest	Light forest																														
Heavy forest	<missing>	Heavy forest	Light forest																														
Heavy forest	Heavy forest	Heavy forest	Heavy forest																														
$H$ matrix for the composite estimator.			<table border="1" style="border-collapse: collapse; width: 100%;"> <tr><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>0</td><td>1</td><td>1</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td></tr> </table>	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1
1	0	0	0	0	0																												
0	1	1	0	0	0																												
0	0	0	1	0	0																												
0	0	0	0	1	0																												
0	0	0	0	0	1																												

Let  $Y_1$  and  $Y_2$  be two compositions (unit-sum vectors) with corresponding covariance matrices  $\Sigma_1$  and  $\Sigma_2$  respectively. The two compositions can have a different dimensionality. Let  $Y_1$  have two classifications (Satellite and Aerial) and let  $Y_2$  have three dimensions. We can also introduce missing data with weighted least squares. Hence, we will let  $Y_2$  have two missing categories. The categories of these compositions are shown on the left side of Table 3.

Let the design matrix  $X$  be specified by the binary matrix depicted in Table 3. This matrix is easily formed by placing a 1 in each cell where all the *available* categories from the original data (rows) match up with their corresponding regression categories (columns).

If we consider  $Y_1$  and  $Y_2$  as partitions of a single vector  $Y$ , and define

$$Y = \begin{matrix} Y_1 \\ Y_2 \end{matrix} \quad \text{and} \quad \Sigma = \begin{matrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{matrix} \quad [5]$$

The generalized regression equation for the new WLS estimate ( $W$ ) and its covariance matrix can be given by

$$W = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y$$

and [6]

$$\Sigma_w = (X^T \Sigma^{-1} X)^{-1}$$

where  $A^{-}$  is the *generalized inverse*<sup>2</sup> of  $A$ .

## DISCUSSION

We have shown that the composite estimator and weighted least squares are both capable of combining multi-level composition estimates of different dimensionality. Both methods require a binary transformation matrix which maps the categories of the corresponding compositions to those of the resulting estimate.

Weighted least squares has an advantage over the composite estimator when multiple estimates need to be combined. It is also unique in its ability to handle data with missing categories. The composite estimator, however, is able to work with a composition estimate of zero covariance. (An example of this is an entire census of a population, where the population statistic represents no sampling error.) In addition, the composite estimator is the only method which can be modified to assure that no components of the resulting estimate are negative.

### Inherent Difficulties with Compositional Data

The unit-sum constraint is an intrinsic property of any composition. Little, if any, attention has been paid to this restriction despite the fact that it greatly complicates any analysis of correlation between categories. Karl Pearson (1897)

---

<sup>2</sup>By definition,  $A^{-}$  is the generalized inverse of  $A$  if:  $A = A A^{-} A$  and  $A^{-} = A^{-} A A^{-}$ . See Graybill (1976).

first pointed out difficulties inherent with the interpretation of correlations between ratios whose numerators and denominators contain common parts.

Since that time, difficulties which have been encountered while trying to interpret these correlations have been described in papers by Chayes (1960,1962), Krumbein (1962), Moisimann (1962), and in books by Chayes (1971) and Le Maître (1982).

Aitchison (1986) points out four basic difficulties which arise from the use of a standard covariance matrix in analysis.

### Negative bias difficulty

For a  $D$ -part composition, we have the restriction

$$\text{cov}(x_1, x_2) + \text{cov}(x_1, x_3) + \cdots + \text{cov}(x_1, x_D) = -\text{var}(x_1) \quad [7]$$

Although the correlations between all parts of a composition would be expected to freely span the range  $[0,1]$  this restriction requires that at least one of the correlations assumes a negative value. Such forced negative correlation could cause some misinterpretation of the data.

### Subcomposition difficulty

In most multivariate situations, an  $m$ -part subset of an  $n$ -part composition would be expected to preserve the covariance structure. For example, consider a 4-part composition of categories A, B, C and D, and then a smaller subset of just the categories A, B, and C.

Table 4.—Correlation matrices from a 4-category sample and its 3-category subsample.

Correlation for {A,B,C,D}					Correlation for {A,B,C}			
	A	B	C	D		A	B	C
A	1.000	0.267	-0.410	-0.484	A	1.000	-0.066	-0.701
B	0.267	1.000	-0.266	-0.419	B	-0.066	1.000	-0.665
C	-0.410	-0.266	1.000	-0.490	C	-0.701	-0.665	1.000
D	-0.484	-0.419	-0.490	1.000				

The above data comes from a hypothetical data set. The correlation matrix on the left side is taken by calculating all four categories. If we instead consider the composition of only the first three categories, the resulting correlation matrix is given on the right side. Notice the correlation between categories  $A$  and  $B$  is positive (0.267), but when removing the fourth category, the same correlation between  $A$  and  $B$  becomes negative (-0.066).

### Basis difficulty

The above correlation matrices were computed from a multivariate sample of unit-sum vectors. When we consider the original data which might be measured (before the unit-sum normalization is performed), we might expect the sample

correlation matrix of the original data to closely resemble the sample correlation matrix of the normalized data. This is not the case, and the two correlation matrices may be dramatically different. This suggests obvious problems when we try to use the covariance or correlation matrices of compositional data to interpret the relationship between two variables.

### **Null correlation difficulty**

It is traditional to consider null-correlation a good indicator of independence. However, in the situations we've already pointed out, a truly independent sample will necessarily have *negative* correlations between elements. It is possible to calculate the naturally occurring negative correlations which would occur under an independence assumption and use this to compare with experimental results, but this problem is also fraught with difficulty.

### **The Contemporary Analogs to a Composite Estimator or WLS**

In his text, Aitchison (1986) provides an approach for analyzing the correlation structure of compositional data using a log-ratio approach. The natural progression for this field of complex sampling designs is to attempt to replace the analysis of covariance matrices with these more contemporary methods. Whether it is possible (and valid) to continue to use the weighted least squares and composite estimator, applying the techniques of Aitchison to approximations derived from the resulting covariance matrices, or if these methods might need to be completely replaced by new methods is currently unknown.

### **ACAS Software Application**

The procedures described in this paper have been developed into a computer package titled "ACAS." (Williams 1995) This application has been written in C++ and is available to the public. Although ACAS is a useable, full-featured package, continual development is underway to provide more advanced tools. The ACAS system already includes some of the methodology of Aitchison and aspires to further develop contemporary methods based on his work.

### **ACKNOWLEDGEMENTS**

The development of this material would not be possible without the ACAS software project, which has been mostly written by David J. C. Beach. This work is currently funded by the State of Minnesota Department of Natural Resources and the USDA Forest Service.

## REFERENCES

- Aitchison, J. 1986. The statistical analysis of compositional data. Chapman and Hall. 416 p.
- Chayes, F. 1960. On correlation between variables of constant sum. *Journal Geophysical Research*, 65:4185-4193.
- Chayes, F. 1962. Numerical correlation and petrographic variation. *Journal of Mathematical Geology*, 70:440-452.
- Chayes, F. 1971. *Ratio Correlation*. University of Chicago Press.
- Czaplewski, R.L. 1992. Accuracy assessment of remotely sensed classifications with multi-phase sampling and the multivariate composite estimator. Hamilton, New Zealand. 16th International Biometric Conference. 2:22.
- Czaplewski, R.L. 1996. Assessment of classification accuracy and extent estimates for a land cover map with double sample. Submitted to *Forest Science*.
- Graybill, F.A. 1976. *Theory and application of the linear model*. Wadsworth & Brooks/Cole Adv. Books and Software. 704p.
- Grizzle, J.E., Starmer, C.F., and Koch, G.G. 1969. Analysis of categorical data by linear models. *Biometrics* 25:489-504.
- Krumbein, W.C. 1962. Open and closed number systems stratigraphic mapping. *Bull. Amer. Assoc. Petrol. Geologists*, 46: 2229-2245.
- Le Maître, R.W. 1982. *Numerical Petrography*. Amsterdam: Elsevier.
- Lee, E.S., Forthofer, R.N., and Lorimor, R.J. 1989. Analyzing complex survey data. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-071. Beverly Hills: Sage Pubns. 78p.
- Mosimann, J.E. 1962. On the compound multinomial distribution, the multivariate  $\beta$ -distribution and correlations among proportions. *Biometrika*, 49:65-82.
- Pearson, K. 1897. Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proc. R. Soc.*, 60:489-498.
- Williams, M.T., and Beach, D.J.C. 1995. ACAS 0.4: accuracy assessment system program manual. USDA Forest Service, Rocky Mountain Forest and Range Experiment Station, Fort Collins, CO. 33 p. + source code.

## BIOGRAPHICAL SKETCH

Murray Todd Williams is currently finishing his M.S. degree in the department of Statistics at Colorado State University. The work which appears in this paper, along with the development of the ACAS software application, is a part of his thesis. Murray received his B.A. in Mathematics from Pomona College in Claremont, CA.

# Cost-Effective, Practical Sampling Strategies for Accuracy Assessment of Large-Area Thematic Maps

Stephen V. Stehman<sup>1</sup>

---

**Abstract.**— Accuracy assessment is an expensive but necessary process in the development and eventual use of a large-scale thematic map. The sampling strategy used for collecting and analyzing the reference samples required for the accuracy assessment must be cost-effective, yet still achieve satisfactory precision for the estimated accuracy parameters. Strata and clusters may be used to improve the efficiency of the sampling strategy, and more specialized designs such as double sampling or adaptive cluster sampling may provide better precision for certain objectives. Post-stratified and regression estimators may be combined with a simple design to yield enhanced precision without substantially increasing costs. Each strategy has strengths and weaknesses and no single strategy is ideal for all applications.

---

## INTRODUCTION

Large-area, satellite-based land cover mapping projects are becoming increasingly important for use in environmental monitoring and modeling, resource inventories, and global change research. Assessing the thematic accuracy of these maps requires balancing the needs of statistical validity and rigor with the practical constraints and limited resources available for the task. Finding this balance is a characteristic of practically any applied sampling problem. The primary expense is obtaining ground reference samples, the critical data for an accuracy assessment. A cost-effective, practical accuracy assessment strategy must efficiently use these ground reference samples. This efficiency can be achieved by selecting an efficient sampling design, or by using efficient analysis procedures following collection of the data.

A site-specific, thematic accuracy assessment is assumed of primary interest, with the evaluation unit being a single pixel as suggested by Janssen and van der Wel (1994). Analogous strategies could be developed for a polygon-based accuracy assessment, but that topic is not pursued here. Each pixel on the target map corresponds to a particular location on the ground, and it is assumed that a ground visit leads to a correct land-cover classification of that pixel. Location error will be confounded with classification error if field site locations do not correspond exactly with map locations. This view of accuracy assessment permits illustration of the basic principles and techniques to be presented in this paper.

The statistical approach taken is the classical finite population sampling model (Cochran 1977). In this perspective as applied to accuracy assessment (Stehman 1995), the population consists of the  $N$  pixels on the map. If a complete census of ground reference sites were obtained, each pixel on the map could be labeled as correctly or

---

<sup>1</sup>*SUNY College of Environmental Science and Forestry, 320 Bray Hall, Syracuse, NY*

incorrectly classified, and a population error matrix constructed from these results in which the rows of the error matrix represent the map classifications, and the columns represent the reference classifications (“truth”). The objective of the accuracy assessment is to use a sample of reference locations to construct a sample error matrix and then to compute estimates of various accuracy parameters, typically the overall proportion of pixels correctly classified ( $P_c$ ), user’s accuracy, producer’s accuracy, and the kappa coefficient of agreement ( $\kappa$ ).

A basic recommendation is that a probability sampling design should be used to collect the reference data. For a probability sample, each pixel in the population has a positive, known probability of being included in the sample. This “inclusion probability” is a characteristic of the sampling design. By knowing the inclusion probability of each sampled pixel, the pixels can be correctly weighted when estimating accuracy parameters. All pixels need not have the same inclusion probability, a stratified sample being a good example, as long as the inclusion probability is known. Probability sampling provides an objective, scientifically defensible protocol for accuracy assessment.

In the finite sampling perspective, a sampling strategy consists of two parts, the sampling design, which is a protocol by which the sample data are collected, and an estimator, which is a formula for estimating a population parameter. An “efficient” sampling strategy is one in which the variance of an estimator is low, or, equivalently, precision is high. Efficiency translates into cost savings in that if a sampling strategy can provide the same variance using fewer sampling locations than another strategy, the former strategy is more cost-effective. Efficiency can be gained by judicious choice of design or estimator, or both. Stratification, cluster sampling, double sampling, and adaptive cluster sampling are design options for improving efficiency. Regression and poststratified estimators are estimation techniques commonly employed in finite population sampling to improve precision. These estimation techniques require auxiliary information in addition to the sample reference data. Aerial photography or videography are potential sources of auxiliary information, and the number of pixels classified into each cover type by the land-cover map can be used as auxiliary information in poststratification.

## ENHANCING PRECISION VIA SAMPLING DESIGN

### Stratification

Several stratification options exist. Stratifying by geographic region ensures that the sampling design for accuracy assessment results in a regionally well-distributed sample, spreads the workload out geographically, and allows for convenient reporting of accuracy statistics by region. For parameters summarizing the full error matrix, such as  $P_c$  and  $\kappa$ , geographic stratification normally does not produce large gains in precision over simple random sampling (SRS) (Cochran 1977, p. 102) and should not be selected with the hope of achieving a substantial reduction in variance for these estimators.

Although geographic stratification is convenient for reporting accuracy statistics by region, it is not necessary to stratify to obtain regional estimates. Any geographic region can be identified as a subpopulation and accuracy statistics calculated for that subpopulation. The advantage of stratification is that the sample size can be controlled in each region or stratum, and the stratum-specific estimates will have smaller variance compared to the estimates obtained via SRS or another equal probability design such as systematic or cluster sampling. For any equal probability design, the sample size in a

region will be proportional to the area of that region, and for small regions, this sample size may be too small to achieve adequate precision.

Stratifying by the land-cover classes identified on the target map has the advantage of guaranteeing a specified sample size in each land-cover class, and thus is cost-effective if the primary objective is estimating user's accuracy. Efficiency for estimating producer's accuracy,  $P_c$ , and  $\kappa$  may be reduced when the design is stratified by the classes identified on the map depending on the allocation of samples to the strata. Stratifying by the map classification also requires that the map be completed prior to selecting the sample, and this may result in a delay between the time of imagery and the time the ground sampling takes place. Changes in land cover may occur during the intervening period.

Still another possible stratification option was employed by Edwards et al. (1996) in a large-area accuracy assessment in Utah, USA. Within broad geographical strata, two additional spatial strata were created, a stratum identified by a 1-km wide corridor centered on a road, and an off-road stratum. The advantage of this stratification scheme is that access to reference sites is much easier for areas in close proximity to a road, and the number of samples per unit cost can be increased dramatically by increasing the sampling intensity in the road stratum. Selecting some samples from the non-road stratum maintains the probability sampling characteristic of the design.

### Cluster Sampling

Cluster sampling is another potentially cost-effective design for selecting reference samples. In this design, the primary sampling unit (*psu*) consists of a cluster of pixels, such as a 3x3 block or a linear row of pixels. In one-stage cluster sampling, all pixels within the *psu* are sampled, whereas in two-stage cluster sampling, a subsample of pixels within each *psu* is selected. The advantage of cluster sampling is that the number of pixels sampled per unit cost is increased because pixels are sampled in closer proximity. Moisen et al. (1994) demonstrated the efficiency gains achievable by cluster sampling based on an analysis taking into account sampling costs and the spatial autocorrelation of classification errors. The pixels within a given sampled *psu* cannot be regarded as independent observations, so the standard error formulas must reflect the cluster structure of the sampling design (Czaplewski 1994, Moisen et al. 1994, Stehman 1996b). The usual SRS variance estimators do not take into account the within cluster correlation and will likely underestimate the cluster sampling variance.

### Sampling Rare Land-Cover Classes

The emphasis placed on sampling rare land-cover categories strongly influences the design selected for accuracy assessment. If rare classes are considered extremely important, the sampling design should reflect this priority and practically forces using a stratified design to ensure adequate sample sizes for these rare classes. At the other extreme, rare classes may be assigned low priority, and an equal probability design is a viable option. Rare classes will not be sampled in large numbers in such designs. An intermediate emphasis on rare land-cover types requires a compromise design. Two general strategies are offered.

The first strategy is to treat rare classes as a separate sampling problem. For example, a general purpose sampling design for accuracy assessment could be a simple random, systematic, or cluster sample. Because such designs will result in only a few reference samples in the rare classes, the general design is supplemented by a specialized

design tailored to sample rare classes with high probability. This two-step design must still be conducted according to a probability sampling protocol, and the analysis must take into account that those pixels selected in the supplemental design may have different inclusion probabilities than those pixels selected by the original design.

Adaptive cluster sampling (Thompson 1990) is tailored to sample efficiently rare, but spatially clustered items. An example illustrating this strategy in an accuracy assessment setting is as follows. Suppose the rare land-cover class is marsh, and assume that a 5x5 block of pixels is adopted as the *psu*. A simple random or systematic sample of *psus* is selected. For those *psus* in which at least one marsh pixel (according to the reference classification) is found, the sampling procedure is “adapted” to then sample adjacent 5x5 blocks surrounding the initial sampled *psu*. If one or more of these adjacent *psus* is found to have at least one marsh pixel, the adaptive strategy is continued. The process stops when no marsh pixels are found in adjacent *psus*. If the rare cover-type is spatially clustered, this strategy will greatly improve the probability of sampling reference pixels in this class because the method intensifies sampling effort in those areas in which the rare class is found. The adaptive design satisfies the necessary criteria of a probability sample. Special estimators of the accuracy parameters need to be employed when this design is used (see Thompson (1990) for the basic theory). Adaptive cluster sampling may be an effective design for change detection accuracy assessment because while change pixels may be rare, they are likely to be spatially clustered, and the adaptive design may more cost-effectively sample such pixels.

Double sampling is another design alternative that may be used to enhance estimation for rare cover types. In double sampling, a large first-phase sample is selected and the stratum to which each sampled pixel belongs is identified. From these first-phase sample units, a stratified random sample, called the second-phase sample, is then selected. The primary advantage of this method is that only the first-phase sample units, not all  $N$  pixels, need to be assigned to strata. The stratification could be based on the land-cover classes as identified by the reference data instead of stratifying on the land-cover classes identified by the map. This stratification is then used to increase the probability of sampling rare ground classes. To implement this design, a large first-phase sample is selected and each location is assigned to a land-cover class. These assignments would not necessarily require a ground visit if reasonably accurate stratum identifications can be made using available maps, aerial photographs or videography. The second-phase sample is then a stratified random subsample of the first-phase sample, and the “true” classification of these second-phase sample sites is identified.

To illustrate how this approach focuses sampling effort on rare categories, suppose a particular class, say marsh, represents only 0.1% of the land area, and that commission error for marsh is high. If the stratification is based on land-cover as identified by the map, it is likely that only a few true marsh sites will appear in the reference sample. By using double sampling and stratifying on the ground classification, the probability of sampling marsh reference pixels could be increased dramatically by intensifying sampling effort in the marsh stratum. Increasing the sample size in this manner will improve precision of producer’s accuracy estimates for the marsh class.

The second strategy for sampling rare classes is to adapt the general purpose sampling design itself to emphasize more strongly rare classes. A stratified design is a typical approach for such a one-step strategy. However, if the general design is chosen to emphasize sampling rare classes, some precision will be lost for other estimates of map accuracy such as  $\hat{P}_c$  and  $\hat{\kappa}$  because of this allocation of sampling resources. This

precision trade-off is characteristic of any sampling design choice, and the properties of the strategy selected should reflect the objectives specified for the accuracy assessment.

## ENHANCING PRECISION VIA ANALYSIS

The second component of a cost-effective sampling strategy is the analysis of the data, provided by the estimators of map accuracy parameters. Three techniques for improving the precision of map accuracy estimates at the analysis stage are poststratification, regression estimation, and incorporating "found" data into an accuracy assessment. All three approaches use auxiliary information in the analysis to achieve this precision gain. Poststratification and regression estimation can be applied to practically any sampling design. Here it is assumed that SRS is employed because an advantage of these analysis techniques is that they can improve precision for simple, easily implemented sampling designs.

### Poststratification

Poststratification is an estimation technique, not a sampling design, requiring auxiliary information similar to that required for stratified sampling. Suppose  $n$  sample pixels are obtained via SRS. Let  $N_{k+}$  and  $n_{k+}$  denote the number of pixels classified as cover type  $k$  in the population and sample, respectively. Both  $N_{k+}$  and  $n_{k+}$  are known once the map is completed. Poststratification incorporates the known totals  $N_{k+}$  into the estimator by analyzing the SRS data as a stratified random sample in which  $n_{k+}$  pixels have been selected from the  $N_{k+}$  pixels available in that stratum (cover type  $k$  as identified on the map). Consider the estimator of  $P_c$ . If  $n_{kk}$  is the number of sample pixels correctly classified in cover type  $k$ , the usual SRS estimator for  $P_c$  is  $\sum n_{kk}/n$ , where summation is over the  $q$  cover types. For SRS, each pixel is weighted equally, the weights being  $N/n$ . For the poststratified estimator, the weights are dependent on the identified strata, the weight for stratum  $k$  being  $N_{k+}/n_{k+}$ . Then the poststratified estimator of  $P_c$  is  $(1/N) \sum_{k=1}^q (N_{k+}/n_{k+}) n_{kk}$ . Poststratified estimators for other parameters are constructed in essentially the same manner, replacing the weight  $N/n$  used in the usual SRS formula by  $N_{k+}/n_{k+}$ , and then summing over the  $q$  strata. Card (1982) presents poststratified estimators and variance estimators for some accuracy parameters, and Stehman (1995) shows the poststratified estimators of  $P_c$  and  $\kappa$ . The precision achieved by poststratification is approximately that of a stratified sample with proportional allocation (Cochran 1977, p. 134), so poststratification will usually result in some gain in precision over the usual SRS estimators. Based on a small simulation study (Stehman 1996c), the gain in precision from poststratification can be expected to be around 5% (in terms of standard error) for estimating  $P_c$  and  $\kappa$ , and as much as 15-30% for estimating producer's accuracy.

A disadvantage of poststratification relative to a stratified design is that the sample size within each stratum,  $n_{k+}$ , is under control with a stratified design, but not with SRS followed by poststratification. However, poststratification can be applied to any classification scheme. For example, if a user wishes to collapse land-cover classes or use an entirely different classification scheme, poststratification can still be applied. Poststratification is flexible because it can be adapted to any identified subpopulation, whereas the usual stratified sampling design is advantageous only for those subpopulations identified as strata prior to sampling. Poststratification and stratified sampling both require  $N_{k+}$ , but these are available once the land-cover map is completed. Because poststratification does not require this information to obtain the sample, no time delay between the imagery and the ground sampling need occur.

## Regression Estimator

In the regression estimator approach, any auxiliary information available that can provide a land-cover classification of a pixel may be used. Primary sources of auxiliary data include aerial photography or videography, or even another land-cover map obtained via remote sensing and using a coarser scale of resolution (e.g., AVHRR). This auxiliary information is separate from the imagery data used to construct the target land-cover map being assessed. A sample of ground reference locations is still required, and these reference data are combined with the auxiliary data via a regression estimator to estimate  $P_c$ . For SRS of  $n$  pixels, the sample proportion of pixels in which the reference and map classifications match is  $\hat{p}_y$ . If no auxiliary data are available,  $\hat{p}_y$  is the usual estimator of  $P_c$ . The sample proportion of pixels in which the map classification matches the classification obtained by the auxiliary data is  $\hat{p}_x$ , and  $\hat{p}_{xy}$  is the sample proportion of pixels in which both the reference and auxiliary data classifications match the map classification. The regression estimator of  $P_c$  is

$$\hat{p}_{reg} = \hat{p}_y + \frac{\hat{p}_{xy} - \hat{p}_y \hat{p}_x}{\hat{p}_x(1 - \hat{p}_x)}(P_x - \hat{p}_x), \quad [1]$$

where  $P_x$  is the proportion of pixels in the entire map in which the auxiliary data and map classifications match. The estimated variance of  $\hat{p}_{reg}$  is

$$\text{var}(\hat{p}_{reg}) = \frac{1 - n/N}{n - 2} \left[ \hat{p}_y(1 - \hat{p}_y) - \frac{(\hat{p}_{xy} - \hat{p}_x \hat{p}_y)^2}{\hat{p}_x(1 - \hat{p}_x)} \right]. \quad [2]$$

The regression estimator results in some gain in precision over the usual SRS estimator, no matter how poor the auxiliary data classifications, but the approach is not worth implementing unless a meaningful gain in precision is achieved. The more accurate the classifications from the auxiliary data, the greater the gain in precision achieved by the regression estimator relative to the usual SRS estimator.

The usual regression estimator approach requires auxiliary information for the entire map region. In practice, it is more feasible to use a double sampling approach in which the auxiliary data are collected for a first-phase sample, and ground visits are made to a subsample of the first-phase sample. The double sampling regression estimator incorporates both the first- and second-phase sample data. Stehman (1996a) demonstrated that the precision of the regression estimator combined with double sampling was nearly the same as the strategy employing the regression estimator with SRS. The double sampling strategy is more cost-effective because it requires obtaining the auxiliary data only on the first-phase sample, not the entire target region.

## Found Data

Additional potential ground reference data may be available via purposeful, haphazard, convenience, or other non-probability sampling methods. For example, suppose an organization has land-cover classifications for various special interest sites which were visited "to see what was there." Assuming the land-cover classifications made at these sites are correct and consistent with the classification scheme employed for the map being evaluated, is it possible to use these data in the accuracy assessment? Such data in a sense represent "free" reference samples because they already exist. But incorporating them into the accuracy assessment in a statistically valid manner is not simple because it is difficult to generalize from these data to the population at large. That is, what population do these "found" sites represent? To illustrate the nature of the difficulty, if reference samples were obtained only from areas within, say, 250 meters

of a road, the sample data can be generalized only to a population of the area in the study region within 250 meters of a road. Generalizing beyond that area is not supported by statistical inferences, and must rest on non-statistical arguments regarding what these sites represent. The same can be said of "found" sites. Because these sites were not selected by probability sampling methods, the population represented by these sites is unknown. The probability sampling protocol must be replaced by assumptions concerning the population represented by the found data. Overton (1990) and Overton et al. (1993) describe some statistical procedures for using found data that apply to this problem. However, the amount of work (cost) needed to use these data in a statistically valid manner appears high, and unless these data are abundant or exceedingly valuable, such as for a very rare cover-type, it is doubtful that such data can contribute significantly to reducing the costs of a statistically rigorous accuracy assessment. Incorporating existing data is much more feasible if a valid sampling design was used to obtain the data.

## GENERAL RECOMMENDATIONS

Some general recommendations for sampling in large-scale accuracy assessment projects are proposed. As with any general recommendation, it is easy to think of situations in which exceptions would obviously be needed. Recommendations should also evolve over time as better methods and new insights are gained. With those caveats in mind, the following suggestions are proposed. For accuracy assessment of large-area land-cover maps, stratification by a few large geographic regions is helpful to ensure a regionally representative sample, to provide adequate sample sizes in these regions for reporting region specific accuracy statistics, and to spread the workload more evenly. As an example of the scale of the recommended stratification, a state such as North Carolina could be stratified into three physiographic regions, the coastal plain, Piedmont, and western upland areas. Bauer et al. (1994) provide another example, as they partitioned their study region (Minnesota) into eight physiographic regions that could be used as strata in an accuracy assessment. States themselves are administratively convenient strata for a land-cover map spanning several states.

Within each geographic stratum, a simple design such as simple random or systematic sampling is recommended as the general design. Implementing a simple design creates the opportunity to employ precision enhancing analysis techniques such as poststratification and regression estimation. Cluster sampling has been demonstrated to be cost-effective (Moisen et al. 1994), and the combination of cluster sampling with stratification into road and non-road areas (Edwards et al. 1996) has considerable practical appeal. Simple designs are easier to implement in the field thus increasing the likelihood that the design will be implemented correctly. This general design strategy is adaptable to a variety of analyses and classification systems, and will accommodate the multiple general uses and objectives present in a large-area mapping project.

To accomplish more specialized objectives, a design tailored to these objectives can supplement the general purpose sampling design. For example, if sampling certain rare classes is an important objective, the data from the general design can be supplemented by an additional simple random sample from each rare class. Users must be careful to incorporate proper weights in the analysis when combining data collected by these two different designs. The double sampling and adaptive cluster sampling designs described earlier may also be used to supplement a simpler, general purpose design.

Accuracy assessment of large-area thematic maps creates several challenging statistical problems. These problems, however, are not insurmountable. The classical

finite sampling approach has been in use for over 50 years and has been applied to sampling large, spatially disperse, and hard to measure populations. National samples obtaining labor, economic, and health statistics are applied to populations over large geographic areas, and these programs face a demanding set of objectives requiring estimates for a variety of parameters at various spatial scales. These problems are characteristic of large-area accuracy assessment efforts. Classical finite sampling theory and methods provide a rich toolbox from which to choose cost-effective sampling strategies, and adapting these strategies to accuracy assessment of large-area thematic maps is a viable approach to pursue.

### ACKNOWLEDGMENTS

I thank Ray Czaplewski for his review and helpful suggestions. This work has been supported by cooperative agreement CR821782 between the U.S. EPA and SUNY-ESF.

### REFERENCES

- Bauer, M.E. et al. 1994. Satellite inventory of Minnesota forest resources. *Photogram. Eng. & Remote Sensing* 60: 287-298.
- Card, D.H. 1982. Using known map category marginal frequencies to improve estimates of thematic map accuracy. *Photogram. Eng. & Remote Sensing* 48: 431-439.
- Cochran, W.G. 1977. *Sampling Techniques* (3rd ed). Wiley: New York.
- Czaplewski, R.L. 1994. Variance approximations for assessments of classification accuracy. Res. Pap. RM-316. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Forest Range and Experiment Station. 29p.
- Edwards, T.C., Jr., Moisen, G.G., and Cutler, D.R. 1996. Assessing map accuracy in an ecoregion-scale cover-map (*in review*).
- Janssen, L.L.F., and van der Wel, F.J.M. 1994. Accuracy assessment of satellite derived land-cover data: A review. *Photogram. Eng. & Remote Sensing* 60: 419-426.
- Moisen, G.G., Edwards, Jr., T.C., and Cutler, D.R. 1994. Spatial sampling to assess classification accuracy of remotely sensed data. In *Environmental Information Management and Analysis: Ecosystem to Global Scales*, W.K. Michener, J.W. Brunt, and S.G. Stafford (eds). New York: Taylor and Francis.
- Overton, W.S. 1990. A strategy for use of found samples in a rigorous monitoring design. *Tech. Rep.* 139, Dept. of Statistics, Oregon State University, Corvallis, OR.
- Overton, J.McC., Young, T.C., and Overton, W.S. 1993. Using 'found' data to augment a probability sample: procedure and case study. *Envir. Monit. and Assmt.* 26: 65-83.
- Stehman, S.V. 1995. Thematic map accuracy assessment from the perspective of finite population sampling, *Inter. J. of Remote Sensing* 16: 589-593.
- Stehman, S.V. 1996a. Use of auxiliary data to improve the precision of estimators of thematic map accuracy (*Remote Sensing of Environment*, in review).
- Stehman, S.V. 1996b. Estimating standard errors of accuracy assessment statistics under cluster sampling (*Remote Sensing of Environment*, in review).
- Stehman, S.V. 1996c. Sampling design and analysis issues for thematic map accuracy assessment. *ASPRS and ACSM Annual Proceedings* (to appear).
- Thompson, S.K. 1990. Adaptive cluster sampling. *J. Amer. Stat. Assoc.* 85: 1050-1059.

### BIOGRAPHICAL SKETCH

Stephen Stehman is an Associate Professor at SUNY-ESF. He has a B.S. in Biology from Penn State, an M.S. in Statistics from Oregon State, and a Ph.D. in Statistics from Cornell. He provides statistical consulting for faculty and graduate students at ESF and teaches courses in sampling, experimental design, and multivariate statistics.

# Spatial and Probabilistic Classification of Forest Structures Using Landsat TM Data

Jeffrey L. Moffett<sup>1</sup> and Julian Besag<sup>2</sup>

---

**Abstract.** Satellite sensors record upwelling radiant flux from the Earth's surface. Classifying forest structures from these measurements is a statistical inference problem, assuming the classes are defined at scales corresponding to the spectral and spatial resolution of the data. Bayesian image analysis methods incorporate both spectral and spatial (contextual) information to improve classification accuracy over spectral classifiers. Stochastic simulation techniques associated with Bayesian models provide a mechanism for estimating class probability vectors for each image pixel. This study evaluates a Markov chain Monte Carlo (MCMC) technique, the Gibbs sampler, for forest classification. The Gibbs sampler runs as a time-homogeneous Markov chain. Class estimates are obtained by seeking the marginal posterior modes (MPM) of the limit distribution. The image is topographically normalized and training samples are collected from randomly distributed pixels. Preliminary results using a simple model show an improvement over maximum likelihood. A hierarchical model designed to incorporate the terrain effects and the system point spread function (PSF) is currently under development.

---

## INTRODUCTION

Inferring forest stand types from image data is a classification process that reduces raw data to useful information. Satellite remote sensing can be useful as a means of stratifying land cover in the first stage of a multi-stage sampling scheme for forest inventory. Classified maps also provide necessary information for landscape planning such as the distribution of stand structure classes by area across a watershed. Probabilistic contextual classifiers are well suited to modelling the uncertainty inherent in the process.

## BAYESIAN IMAGE ANALYSIS METHODOLOGY

For a given scale, the desired classification can be modelled as a discrete Markov random field (MRF)  $x$  with an associated probability distribution  $\pi(x)$ .

---

<sup>1</sup>Graduate student, Silviculture Laboratory, University of Washington, Seattle, WA ([jmoffett@silvae.cfr.washington.edu](mailto:jmoffett@silvae.cfr.washington.edu)).

<sup>2</sup>Professor, Department of Statistics, University of Washington, Seattle, WA.

This probability distribution can be combined with a likelihood density for the data  $y$  given  $x$ . The combination gives a posterior probability distribution for the random field. The posterior distribution is thus a conditional probability distribution for the classification given the image data. Statistical inferences about  $x$  are then based on the posterior distribution  $\pi(x|y)$ .

There are several advantages of this approach. In addition to classifying each pixel on the basis of the spectral information contained in the image data, other *a priori* information can be specifically modelled by the prior distribution. This information pertains to the local characteristics of the classification, *e.g.* having adjacent pixels belonging to the same class, having objects with closed boundaries, or preventing two particular classes from being adjacent. Unlike post-classification smoothing algorithms, such as majority filters, Bayesian classification techniques incorporate spatial information without losing fidelity to the data.

### Observed data

Consider a two-dimensional rectangular array of pixels over a finite region  $S$  with each pixel identified by an index  $i = 1, 2, \dots, n$ . For each pixel in  $S$  a remote sensor records a measurement of radiant flux that is a sample distribution of the upwelling radiance, degraded by a stochastic process. The imperfect observations form an image  $y = \{y_i: i \in S\}$ . The  $y_i$ 's are vector valued for bands 1-5,7.

### Markov random fields

Let  $X$  be a random vector that classifies  $y$  over region  $S$  and  $x = \{x_i: i \in S\}$  be any realization of  $X$ . Thus,  $x$  denotes an arbitrary classification of  $S$  and  $X = \{X_1, \dots, X_n\}$ . Each univariate  $X_i$  takes values among  $c$  unordered classes. The objective of classification is to estimate the "true" class image  $x^*$ . Bayesian image analysis assumes  $x^*$  is a realization of a locally dependent MRF with an associated joint probability distribution  $\pi(x)$  assigning classes to  $S$  (Besag, 1986). In forest type mapping stands are typically larger than the individual pixels. Exploiting this spatial continuity of the surface cover at a particular scale relative to the image resolution provides a second source of imperfect information about  $x^*$ .

Unlike time dependent variables the only logical conditioning set for spatial variables is all  $x$  except  $x_i$ . An MRF must be defined with respect to a set of neighbors, written as  $x_{\partial i}$ , and satisfy the positivity condition:  $\pi(x) > 0$  for all  $x$ . Given a specified neighborhood system, the Markov property states  $\pi(x_i|x_{\cdot i}) \equiv \pi(x_i|x_{\partial i})$ , where  $x_{\cdot i}$  represents all pixels other than pixel  $i$ . The Markov property allows the full conditional distribution of  $x_i$  to be specified in terms of its neighbors. As a result of this assumption,  $\pi(x)$  is a Gibbs distribution with respect to the assumed neighborhood (Geman and Geman, 1984). The MRF-Gibbs equivalence provides an explicit formula for the jointly distributed random field, in terms of the full conditionals or Gibbs potential functions.

A second-order neighborhood consists of eight neighbors in the horizontal, vertical, and diagonal directions. The general form of a pairwise interaction MRF for discrete unordered classes can be written as follows (Besag, 1986):

$$\pi(x) \propto \exp\left( \sum_{1 \leq k \leq c} \alpha_k n_k - \sum_{1 \leq k < l \leq c} \beta_{kl} n_{kl} \right),$$

where  $n_k$  is the number of pixels in class  $k$  and  $n_{kl}$  is the number of distinct class pairs  $(k, l)$ . The  $\alpha_k$ 's and the  $\beta_{kl}$ 's govern the percentage of pixels in each class and the strength of interaction between different classes; for  $\beta_{kl} > 0$  classes  $k$  and  $l$  are discouraged from being neighbors.

### Likelihood function

The recorded spectral radiance influences the classification through the likelihood function. In contrast to the prior, the likelihood is ideally determined by knowledge of the sensing process and the existence of training data. The likelihood function represents the joint probability density as a conditional density of the data  $y$  given the actual scene  $x$ :

$$L(y | x) = \prod_{i \in S} f(y_i | x_i).$$

Written in this form the image data is assumed to be conditionally independent given knowledge of the classes, or class conditionally independent. While this assumption expedites an overview of the theory, and may be reasonable for some applications, it is not a valid assumption for remotely sensed images.

### Posterior distribution

Bayes theorem can be used to construct a posterior distribution  $\pi(x|y)$  on which inferences about  $x^*$  can be based. Using this theorem, the posterior distribution for the scene  $x$  given the data  $y$  can be written as,

$$\pi(x | y) \propto L(y | x) \pi(x).$$

The computational burden presented by the posterior distribution is overcome by the fact that it too is an MRF with a Gibbs distribution. This allows pixels to be evaluated individually based on their full conditional distributions, written as

$$\pi(x_i | x_{-i}, y) \propto f(y_i | x_i) \pi(x_i | x_{\bar{i}}).$$

The conditional prior distribution follows from  $\pi(x) = \pi(x_i | x_{-i}) \pi(x_{-i})$ .

### **Image estimation**

Within this framework the final classification is a point estimate of  $x^*$ . Due to the high dimensionality of  $x$  the optimization required to find an exact solution is infeasible for realistic scenes. Estimates of various image attributes can be obtained using stochastic simulation. The marginal posterior modes (MPM) estimate will minimize the expected number of misclassifications by maximizing the marginal probability with which each pixel is classified. MPM has been increasingly adopted in recent years for classification problems (Besag *et al.*, 1991; Green *et al.*, 1994). MPM estimates of  $x_i^*$  are obtained by maximizing  $\pi(x_i|y)$ , where the most frequently sampled class for pixel  $i$  is the estimate of  $x_i^*$ .

This estimate can be computed using the Gibbs sampler (Geman and Geman, 1984) which runs as a discrete-time Markov chain, with state space the set of all possible classifications  $x$  and limit distribution  $\pi(x|y)$  (Besag, 1989). For a single iteration, a new  $x_i$  is randomly sampled for each pixel from its full conditional distribution. This generates a sequence of stochastically dependent classifications  $\{x\}$ . An initial classification is required to start the algorithm followed by a number of iterations during which samples are not collected. This burn-in period will minimize the influence of the initial classification. This procedure also estimates a posterior probability distribution for each pixel, providing spatially explicit information about the certainty of the point estimate classification. Credible intervals can also be obtained for measurements such as class areas. The MPM estimate requires simulating a predetermined number of realizations of the MRF and it can be difficult to determine the number of realizations required to assure convergence. Parameter estimation can also be problematic.

## **A PRELIMINARY MODEL**

Using this approach a project was undertaken to classify TM data for the Lizard Lake watershed near Clallam Bay, Washington. The area is covered by second growth forests. Six classes were selected to represent general forest structures and land cover: clearcut (CC); stand initiation (SI); stem exclusion (SE); older forest (OF); hardwood (HD); and deep water (W). Class means and covariances were estimated from randomly selected pixels within training sites to avoid biasing the estimates. As this is an area of rugged terrain, the backwards radiance correction transformation (Colby, 1991) was used to topographically normalize the image.

### **The model**

A simple model was constructed by first adopting the conventional maximum likelihood function (ML), which assumes the data to be class conditionally independent and normally distributed. A discrete pairwise difference prior was chosen to model the tendency of adjacent pixels to be of the same class. The prior distribution is written as follows,

$$\pi ( x ) : \pi ( x_i ) \propto e^{\beta n_i(c)},$$

where  $\beta$  is a strength of interaction parameter, and  $n_i(c)$  counts the number of neighbors (8 nearest) of pixel  $i$  in class  $c$  for  $c = \{0, \dots, 5\}$ .

## Results

The Gibbs sampler was written in C. Samples were collected for 3000 iterations after a burn-in of 1000 iterations;  $\beta$  was fixed at 0.5. Figure 1 shows the "ground truth" classification, which does not indicate the presence of hardwood patches known to be on this landscape. Figure 2 shows the ML estimate from data that was not normalized. The topographic shade causes some SE forests to be classified as OF. Figure 3 shows the MPM point estimate from normalized data. Figure 3 also shows a single pixel at the center of a small lake where the water is deepest (lower left corner). The influence of the likelihood preserves the single deep water pixel, which would be lost using a majority filter. Figure 4 shows a posterior probability classification of SE forest. Note that the edges of the younger forest types and older forests have a modest probability of being stem exclusion.

## A HIERARCHICAL MODEL

The likelihood in the above model is problematic and the prior is the most basic. Sources of variation such as atmospheric scattering and topography complicate the classification process. As an alternative to data preprocessing, the Bayesian approach allows separate models for each source of variation to be linked through successive conditioning of the form  $\pi(x|z_1)$ ,  $\pi(z_1|z_2)$ , ...,  $\pi(z_k|y)$  (Smith and Roberts, 1993). Each source of variation is modelled by a conditional probability distribution as a stage within a hierarchical framework. If a second-order neighborhood is assumed, triple and quadruple pixel interaction can be modelled (Besag, 1974). Triples are useful for modelling convex stand boundaries and quadruples for modelling interior regions. Recent research suggests that for restoration problems a crude prior will suffice, but that estimates of attributes such as class area and perimeter are sensitive to the specification of the prior distribution (Tjelmeland and Besag, 1996).

### Topographic effect model

Estimates of slope and aspect are generally calculated from digital elevation models (DEMs), but are subject to several sources of error. Given the likely errors and little knowledge of the surface photometric function it is therefore logical to model the effect of topography with some degree of uncertainty. This can be specified by first defining  $z$  to be the measured pixel image incorporating the effects of topography. Assuming  $z$  to be conditionally independent given  $x$ , and each  $z_j$  to be normally distributed gives the following model,

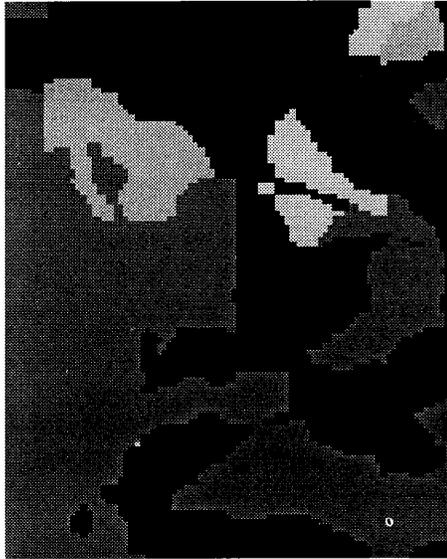


Figure 1. "Ground truth." Note missing patches of hardwoods.



Figure 2. ML classification without topographic normalization showing a greater portion of older forest.

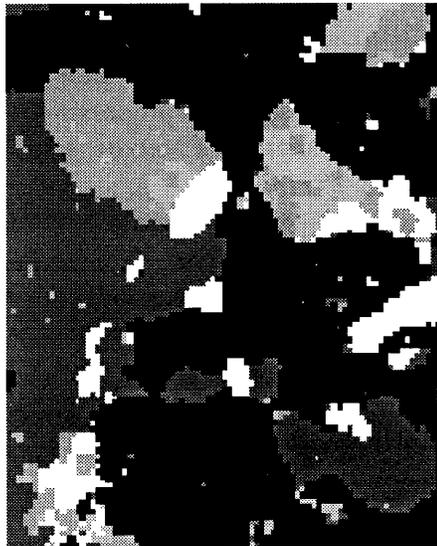


Figure 3. MPM point estimate,  $\beta=0.5$ .

■	W	▨	SI
□	HD	■	SE
▨	CC	■	OF

Legend for Figs. 1,2,3.



Figure 4. Posterior probability of SE forest,  $\beta=0.5$ . Linear stretch with brighter shades indicating higher probability.

$$\pi( z | x, DEM, sun\ position) \propto \exp(-\frac{\theta}{2} \sum_{j=1}^n (z_j - (\frac{\cos^k i \cos^k e}{\cos e}) \mu_{x_j})^2),$$

where  $\cos i$  is the cosine of the angle of incidence,  $\cos e$  is the cosine of the slope,  $\theta$  is the variance,  $\mu_x$  are the means, and  $k$  is the Minnaert constant. This model is derived from the normalization transformation of Colby (1991). The assumption of normality in this model is arbitrary, requiring further investigation.

### System PSF effects model

After reflecting off of a topographic surface the upwelling radiant flux is scattered by the atmosphere and recorded as a function of the sensor PSF (Duggin and Robinove, 1990). If the data  $y$  are assumed normally distributed and conditionally independent given  $z$ , then this process can be modelled as follows,

$$\pi( y | z ) \propto \exp(-\frac{\phi}{2} \sum_{i=1}^n (y_i - \sum_j h_{ij} z_j)^2),$$

where  $h$  is the system PSF detecting radiation from  $j$  pixels in the proximity of pixel  $i$ , and  $\phi$  is the variance. Note also that  $h$  will be a function of wavelength.

### Posterior distribution

Bayes theorem allows these models to be combined and gives the following joint posterior distribution,

$$\pi( x, z | y ) \propto \pi( y | z ) \pi( z | x ) \pi( x ).$$

Again an MPM estimate can be obtained by sampling from the full conditionals. However the added complexity of this model may require the use of a Metropolis algorithm. The complexity will be further increased if prior distributions are specified for some of the parameters as a fully Bayesian approach would suggest.

## CONCLUSION

Appropriate accuracy assessment methods for probabilistic classifications must still be developed. Estimates of posterior probabilities will be difficult to verify. The "ground truth" for assessing point estimates needs to be at the same scale and resolution as the image data. Sensitivity analysis must be conducted to assess the stability of the solutions. Topographic normalization and spatial information can improve forest classification. Probabilistic classifiers provide additional information by estimating the level of certainty with which a pixel is classified,

which is the logical result of an inference procedure.

### ACKNOWLEDGEMENTS

Many thanks to C. Oliver, G. Klock, J. McCarter, G. Schreuder, and M. Smith.

### REFERENCES

- Besag, Julian. 1974. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society, Series B.* 36(2):192-236.
- Besag, Julian. 1986. Statistical Analysis of Dirty Pictures. *Journal of the Royal Statistical Society, Series B.* 48:259-302.
- Besag, J. 1989. Towards Bayesian Image Analysis. *J. of Appl. Stat.* 16:395-407.
- Besag, J., J. York, and A. Mollié. 1991. Bayesian Image Restoration with Two Applications in Spatial Statistics. *Ann. of the Ins. of Stat. Math.* 43(1):1-59.
- Colby, Jeffrey D. 1991. Topographic Normalization in Rugged Terrain. *Photogrammetric Engineering and Remote Sensing.* 57(5):531-537.
- Duggin, M. and C. Robinove. 1990. Assumptions Implicit in Remote Sensing Data Acquisition and Analysis. *Int. J. of Remote Sensing.* 11(10):1669-1694.
- Geman, S. and D. Geman. 1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. PAMI.* 6(6):721-741.
- Green, E.J. *et al.* 1994. Construction of Thematic Maps from Satellite Imagery. In: *Bayesian Statistics 5. Proceedings of the 5th Valencia Int. Meeting.* p. 75.
- Smith, A. and G. Roberts. 1993. Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. *J. R. Stat. Soc., B.* 55(1):3-23.
- Tjelmeland, Håkon and Julian Besag. 1996. Markov Random Fields with Higher Order Interactions. submitted for publication.

### BIOGRAPHICAL SKETCH

Jeffrey Moffett is researching MCMC classification of RS imagery for a PhD dissertation and provides GIS support for the Landscape Management Project.

Julian Besag works mainly in spatial statistics and Bayesian inference. He is the Director of the Center for Spatial Statistics at the University of Washington.

# Total Error Estimation in a Spatial Database for GIS

Jose Alberto Quintanilha<sup>1</sup> and Marcos Rodrigues<sup>2</sup>

---

**Abstract.** - The aim of this paper is to present a methodology to estimate the errors in digital spatial database originated from the digitizing a large number of maps. It is understood that the desired quality of that digital database is associated to a specific foreseen use. Supposing that the spatial database is the population to be sampled and the objects on it as the primary sampling units, two strategies are discussed: the spatial, focused on areas where errors are likely to occur and the non-spatial sampling, focused on the selection of objects according to a database structure. The errors types are organized in a sequential structure upon which a multi-stage sampling strategy is applied. The sampling areas are selected by a specific sampling frame, quadrats, selected according to spatial variability evaluated by entropy measures. Supposing that the amount of errors in the spatial units have a Poisson distribution, the total error distribution can be estimated, allowing thus the inference of the whole population quality.

---

## INTRODUCTION

The purpose of this paper is to present a methodology to estimate the total errors in a spatial database for GIS applications. The population to be sampled is the database from the digitalization of a large amount of maps in paper format.

The adopted sampling schema involve spatial and non-spatial aspects. The spatial aspects concern the selection of spatial areas where errors are more likely to occur; the non-spatial aspects are related to the selection of spatial objects (points, lines and polygons) structured in a particular data model.

The sampling area frame, quadrats, are selected according to their variability, measured by their correspondent entropy.

A sequence of non-independent errors in database: identification (completeness), absolute position, relative position and topology, in that order (for more details, see Quintanilha, 1996), was used under a multiple stage sampling procedure.

The total errors in a spatial database can be estimate under the assumption that the frequencies of the sequence of the errors follow a multivariate Poisson distribution (Johnson and Kotz, 1969).

---

<sup>1</sup> Universidade de Sao Paulo - Brazil

<sup>2</sup> Universidade de Sao Paulo, CARTA Consultoria - Brazil

The aim is to obtain an estimate for each type of error examining the smallest number of maps. A methodology to analyse the frequency of each type of errors (completeness, absolute position, relative position and topology) in a spatial database is proposed.

## **EVALUATION OF THE ERRORS IN A SPATIAL DATABASE**

As mentioned before, the evaluation of the sequence of errors is made by the following steps:

- step 1: extract a sample set of maps (random or intentional) from the different existent sets of maps (non-spatial sampling);
- step 2: extract a sample of maps from the selected set of maps from the first step (non-spatial sampling);
- step 3: choose areas from the selected maps from step 2 (spatial sampling);
- step 4: stratify selected areas by themes or coverage (non-spatial stratification);
- step 5: select the objects from the database by themes or coverages of interest (non-spatial sampling).

In steps 1, 2, 3 and 5, Shannon's entropy concept is used to select the set of maps, the maps and the area to be sampled within each map by themes and by objects. Considering that high variability implies high probability to observe the errors under study, one selects the set of maps, maps, and areas with higher entropy value, which may be estimated by the number of bits of the correspondent datafile. So, sampling spatial database was reduced to locate the spatial sampling units using information theory.

## **SPATIAL SAMPLING**

### **Spatial Sampling Techniques**

We aim at selecting an adequate area to identify and count errors to certify the database accuracy. Therefore all the spatial units and spatial sampling frames to be used are area type.

The most common sampling frame in spatial sampling is the quadrat. A quadrat (Shaw and Wheeler, 1985) is the result of a superposition of a regular grid on the map (or area). Errors are counted within the quadrat.

Kish (1965), affirms that quadrat sampling is not a sampling method, but a methodology to collect information in a spatial position selected by some traditional sampling scheme (random, systematic, stratified, etc.).

The quadrat sampling is recommended when the units of the population are not available (either access to them is difficult or the associated cost is very high). The

case of errors evaluation in database is an example of the inviability of census for reasons of time and cost.

The quadrats sampling can be used in two different purposes: measure the abundance per areal unit or investigate the population pattern (Ripley, 1981). In the first case, one may suppose that the errors in the quadrat follow a Poisson process with tax  $\lambda$  (error rate i.e., number of errors occurrence per area unit). The frequency of errors in a quadrat with area  $A$  will follow a Poisson distribution (mean and variance equal to  $\lambda A$ ). In the second case, indexes have been proposed that are based on counting from a set of samples of quadrats (Ripley, op.cit.).

Formally:

1) if  $N(A)$  is the observed number of errors in area  $A$  and  $P_n(A) = \text{prob} \{N(A) = n\}$ ,  $\exists \lambda > 0$ , constant, implies that for any quadrat with area  $\delta A > 0$  (Diggle, 1979 and 1981):

$$P_n(\delta A) = \begin{array}{ll} 1 - \lambda \delta A + o(\delta A) & (n=0) \\ \lambda \delta A + o(\delta A) & (n=1) \\ o(\delta A) & (n=2,3,\dots) \end{array}$$

where  $o(\delta A)$  denotes magnitude orders of real numbers and vectors sequences.

2) if for any two exclusive and congruent regions,  $A$  and  $B$ ,  $N(A)$  e  $N(B)$  are independents. Then:

$$P_n(A) = e^{(-\lambda A)} (\lambda A)^n / n! \quad (n=0,1,\dots).$$

Thus,  $N(A)$  follows a Poisson distribution with mean equal to  $\lambda A$ .

### **Determination of quadrats size or number of quadrats**

The determination of the number of quadrats are related to a cost model, which involves the precision of the estimates and, obviously, the determination of costs and times. For a certain budget or duration for the database quality investigation, the spatial objects in the set of maps, maps, areas, themes/coverage can be investigate. In others words, for an expect value for the error of the estimatives, we can determine an ideal size of the sample to satisfy this precision.

Certainly, a higher number of small quadrats will cover the area more efficiently. So, the chance to detect some type of error will be improved and the correspondent spatial correlation will be minimized. However, when sampling units types are areas and lines, the definition of the quadrats must consider the problem of the size of these objects which can transcend the area of the quadrat. One example is the segmentation of polygons in meaningless lines.

### Determination of position of the quadrats

For a certain area represented in a quadrat and for a certain representation model, raster or vector, the size of datafile which represents that quadrat is variable, proportional to the number  $c$  of different objects inside it. For each selected area,  $p_i$  is the probability of occurrence of the  $i$ th object, which is estimated by their frequency. The size of the datafile related to a quadrat is proportional to the number  $c$  too, i.e.:

$$\sum_{i=1}^c p_i = 1.00$$

Considering  $t_i$  as the size of a part of a quadrat datafile related to object  $i$ , the expected mean size of the quadrat datafile is given by:

$$E(t) = \sum_{i=1}^c p_i t_i$$

The expected size of the quadrat datafile  $E(t)$  is proportional to the sum of the number of different objects within it.

Similarly, the entropy of the quadrat is given by:

$$H = - \sum_{i=1}^c p_i \log_2 p_i \text{ bits/quadrat.}$$

So, the sampling procedure must be restricted to areas where the entropy is higher, which corresponds to the areas with higher number of bits. The maps and sets of maps will be selected similarly.

The themes/coverage, will be used like a secondary variable in the objects stratification: objects associated with each theme/coverage in areas with the higher quadrat datafile will be in the sample.

### Use of the thematic stratification

The option for a stratified sampling by themes/coverages is justified since the simple random sampling, relates to the global distribution of the population (i.e., the errors), which implies that some type of errors may not be sampled or may be undersampled. This could be avoided with a larger sample but then, there would be implications on the time and associated costs.

## TEST FOR THE TOTAL ERRORS AMOUNT

As mentioned, we use the sequence present above and consider each type of error (identification, absolute position, relative position, and topology) as a Bernouille experiment. That is, there are two possible results for each type of error: it exist or not.

The assumptions are:

- the vector  $\mathbf{Y} = (Y_I, Y_A, Y_R, Y_T)'$  is the vector of the number of errors each type errors in  $n$  quadrats, where:
  - $Y_I$  is the number of errors of identification,
  - $Y_A$  is the number of errors of absolute position which was not generated by errors of identification,
  - $Y_R$  is the number of relative position which was not generated by the absolute position errors,
  - $Y_T$  is the number of errors of topology which was not generated by the relative position errors;
- each type of errors follows a Poisson distribution;
- the four kinds of errors are correlated (identification errors implies absolute position errors which implies relative position errors which implies in errors in topology).

So, the vector  $\mathbf{Y}$  will follow a Poisson multivariate distribution with the parameters:  $\lambda^* = (\lambda_I, \lambda_A, \lambda_R, \lambda_T)$  where  $\lambda_i$  is the tax of error type  $i$ , where  $i = I$  (identification),  $A$  (absolute position),  $R$  (relative position),  $T$  (topology). Details of this distribution can be found in Johnson and Kotz (1969); Jensen (1985) e Ho (1995).

## PROPOSED METHODOLOGY

Consider a situation when a number of maps with the same cartographic characteristics (scale, *datum*, coordinate system, generalization conditions, legends, etc.) is necessary to cover a whole area. Over that maps, one overlays a regular grid which create a similar number of equal quadrats in each map. The overall procedure is:

- create a homogeneous sets of maps,
- select a sample of set of maps according of the entropy criteria,
- for each selected set of maps, extract a sample of maps according the same criteria (entropy),
- for each map in the sample, locate the quadrats with the higher entropy,
- for each theme/coverage in each selected quadrat, compute:
  1. the number of identification errors,

2. the number of absolute position errors not generated by the identification errors;
3. the number of relative position errors not generated by the previous errors;
4. the number of topology errors not generated by the previous errors.

The event error can be represented by:

$$\text{error} = I \cup (\sim I \cap A) \cup (\sim I \cap \sim A \cup R) \cup (\sim I \cap \sim A \cap \sim R \cap T).$$

Because they are exclusive events, the error occurrence probability can be estimated by:

$$\begin{aligned} p(\text{error}) &= p(I) \text{ or } p(\sim I \text{ and } A) \text{ or } p(\sim I \text{ and } \sim A \text{ and } R) \text{ or } p(\sim I \text{ and } \sim A \text{ and } \\ &\quad \sim R \text{ and } T) = \\ &= p(I) + p(\sim I \cap A) + p(\sim I \cap \sim A \cap R) + p(\sim I \cap \sim A \cap \sim R \cap T) = \\ &= p(I) + p(A/\sim I) * p(\sim I) + p(R/\sim A \cap \sim I) * p(\sim A/\sim I) * p(\sim I) + \\ &\quad + p(T/\sim I \cap \sim A \cap \sim R) * p(\sim R/\sim A \cap \sim I) * p(\sim A/\sim I) * p(\sim I) \end{aligned}$$

where:

$p(I)$  is estimated by the number of identification error, i.e., the proportion of non identified objects or objects that have been erroneously identified,

$p(\sim I) = 1 - p(I)$  is estimated by the proportion of objects correctly identified,

$p(A/\sim I)$  is estimated by the proportion of objects with absolute position errors that were not generated from the identification errors,

$p(R/\sim A \cap \sim I)$  is estimated from the proportion of objects with relative position errors that were not generated from the previous errors,

$p(\sim R/\sim A \cap \sim I)$  is estimated from the proportion of objects without identification, absolute position and relative position errors,

$p(T/\sim I \cap \sim A \cap \sim R)$  is estimate from the proportion of objects with topology errors that were not generated from the previous errors.

## CONCLUSIONS

That proposed methodology for errors evaluation in spatial database, aims at reducing the time and costs associated to qualify and validate that spatial database for GIS users. Additional studies will be carried out to handle the spurious polygons that can be generated for each type of errors. These polygons may complicate the sampling procedure and increase the sample size, time and costs.

## ACKNOWLEDGMENTS

The authors are grateful to Escola Politécnica da Universidade de Sao Paulo which supported this research and to Dr. Linda Lee Ho, who supported the statistical aspects of Poisson multivariate distribution.

## REFERENCES

- Bartlett, M S. 1975. The statistical analysis of spatial pattern. Chapman and Hall, London, 90 p.
- Bellhouse, D. R. 1988. Spatial sampling. In: Kotz, S.; Johnson, N.L. (ed.) Encyclopedia of Statistical Science, New York, John Wiley, 1988. V. 8, p. 581-4.
- Berger, T. 1981. Information theory and coding theory. . In: Kotz, S.; Johnson, N.L. (ed.) Encyclopedia of Statistical Science. New York, John Wiley, V.4, p. 125-41.
- Diggle, P. J. 1979. Statistical methods for spatial point patterns in ecology. In: Cormack, R.M.; Ord, J.K., ed. Spatial and temporal analysis in ecology. Fairland, International Co-operative Pub. House, p.95-150.
- Diggle, P. J. 1981. Some graphical methods in the analysis os spatial point patterns. In: Barnett, V., ed., Interpreting multivariate data. Chichester, John Wiley, chap. 4, p. 55-73.
- Ho, L. L. 1995. Analysis of multivariate counts. São Paulo. 109p. Thesis (PhD) - Escola Politécnica, Universidade de São Paulo, Brazil (in portuguese).
- Jensen, D. R. 1985. Multivariate distributions. . In: Kotz, S.; Johnson, N.L. (ed.) Encyclopedia of Statistical Science, New York, John Wiley, V.6, p. 43-54.
- Johnson, N.L.; Kotz, S. 1969. Distributions in statistics: discrete distributions. New York, John Wiley. Chap.11: Multivariate Discrete Distributions, p.281-323.
- Kish, L. 1965. Survey sampling. Chapter 9: Area sampling. p. 301-358. New York, John Wiley.
- Quintanilha, J. A. 1996. Errors and uncertainties in a spatial database for GIS (non-official version). 193p. Thesis (PhD) - Escola Politécnica, Universidade de São Paulo, Brazil (in portuguese).
- Ripley, B. D. 1981. Spatial statistics. New York, John Wiley & Sons, 252 p.
- Ripley, B. D. 1988. Spatial data analysis. . In: Kotz, S.; Johnson, N.L. (ed.) Encyclopedia of Statistical Science. New York, John Wiley, V. 8, p. 570-3
- Shaw, G.; Wheeler, D. 1985. Statistical techniques in geography. S.I.: Spatial indices & pattern analysis, Chap. 16.

# Assessing the Accuracy of a Regional Land Cover Classification

William Clerke<sup>1</sup> Raymond Czaplewski<sup>2</sup>  
Jeff Campbell<sup>3</sup> and Janet Fahringer<sup>3</sup>

## ABSTRACT

---

The Southern Region USDA Forest Service recently completed the Southern Appalachian Assessment (SAA). The Assessment is a broad scale interagency analysis and sharing of existing information relative to the natural and human resources of the region. The SAA encompasses over 36 million acres extending from Northern Virginia to Northern Alabama. It was clear early in the Assessment that a uniform land cover data layer would be required to support analysis and modeling.

Pacific Meridian Resources (PMR), in cooperation with the Forest Service, developed a 16-class land cover map of the Assessment area. The classification included nine Anderson level 3 forest cover classes. Land cover polygons, based on a minimum mapping unit of two acres, were derived from the final pixel classification. This data set, in raster and vector formats, was provided to the Assessment participants as an input for deriving required products. The Forest Service took the lead in cooperation with PMR in conducting an accuracy assessment to validate the polygon land cover data theme.

A sampling design included aerial photo interpretation and ground examination phases was developed. During the aerial photo phase, data was collected on approximately 1500 sample polygons within 239 systematically distributed Primary Sampling Units (PSU's). Sample selection procedures were implemented that accounted for the skewed spatial and proportional distribution of the land cover classes within the Assessment area.

---

<sup>1</sup> Remote Sensing Specialist, Southern Region, USDA Forest Service, Atlanta, GA.

<sup>2</sup> Biometrician, Rocky Mountain Forest and Range Experiment Station, Fort Collins, CO.

<sup>3</sup> Project Manager and Senior Analyst respectively, Pacific Meridian Resources, Portland, OR.

# Sampling Satellite Images for Area Estimates in a Large Region

F.J. Gallego<sup>1</sup>

---

**Abstract.** This paper describes two alternative methods (square sites and strips) to select a sample of sites for satellite image analysis giving priority to a particular type of area (intensive agricultural or forest) when no suitable stratification is available. Both methods are based on two-phase sampling. For square sites, a systematic sample is first drawn on the SPOT reference grid. Pre-selected sites are visually photo-interpreted for a rough estimation of the percentage of agricultural land or forest, that determines the sampling probability in the second phase. The alternative scheme, based on thin satellite image strips, seems to be more efficient than sampling square sites according to a simulation study based on ground survey data in Spain.

---

## INTRODUCTION

We consider here two problems for which high resolution satellite images are sampled because analysing a complete coverage of is too expensive:

- Area estimation of a particular land cover (an annual crop, forest, etc.) or estimation of area change between two dates.
- Accuracy assessment of a land cover map obtained by other means, for example by multi-temporal analysis of low resolution satellite images.

The Institute for Remote Sensing Applications (IRSA) of the European Commission (EC) is involved in several projects where these problems are tackled:

1. Rapid estimates of area change for main crops in the European Union (EU) of the MARS Project (Monitoring Agriculture with Remote Sensing).
2. Multi-country crop area estimates in central Europe. This is part of the MERA-95 Project (MARS and Environmental related Applications).
3. FIRS Project (Forest Information from Remote Sensing). Forest mapping and estimation in Europe, from the Atlantic to the Urals (Kennedy et al 1995).
4. TREES Project (Tropical Ecosystem Environmental observations by Satellite): Forest mapping in the inter-tropical belt (Malingreau et al 1995).
5. Estimation of crop area change in countries of the former Soviet Union.

---

<sup>1</sup> MARS Project, IRSA, JRC, 21020 Ispra (Va), Italy

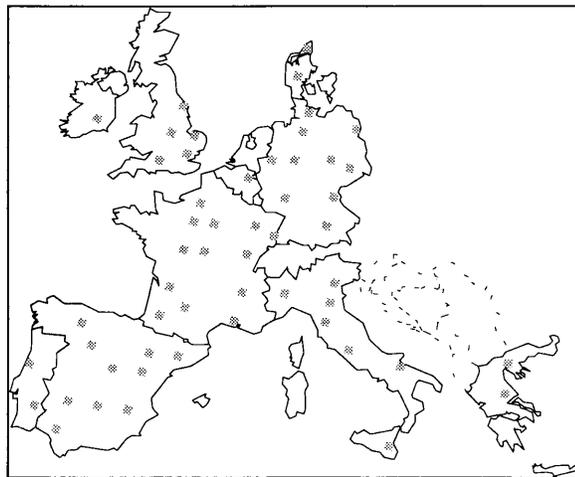
The examples we present below refer to items 1 and 2 (MARS Project), where the purpose is mainly statistical. Mapping is not a priority. For FIRS and TREES (items 3 and 4) mapping is the main priority, but area estimation and evaluation of mapping accuracy are associated targets (Mayaux and Lambin, 1995). Item 5 refers to a contract of the EC with the French company “SCOT Conseil”, in which IRSA will be responsible for site sampling.

### **SAMPLING SITES OF 40 KM × 40 KM**

This unit size is reasonable for multitemporal analysis of high resolution satellite images. They are large enough to include a variety of land cover types in most landscapes, and, if conveniently located, fit inside marketed images.

#### **The “Rapid Crop Area Estimates in the EC” of the MARS Project.**

Rapid crop area estimates are being produced in the European Union by analysis of high resolution satellite images on a sample of sites of 40 km × 40 km (figure 1). Estimates are computed for inter-annual variations, not for a particular year (Pous et al, 1995, Sharman and De Boissezon, 1992). The method is run on a non random sample of 53 square sites of 40 km × 40 km for EU-12 (The 12 members before 1995). With the 3 new member states (Austria, Finland and Sweden), the number of sites has become 60. Ground surveys are conducted on a sample of approximately 16 segments of 1400 m × 1400 m in each site (Carfagna 1995): 40 points per segment are visited and a subsample of farms are interviewed (Gallego et al 1994). Some problems arise to extrapolate results because of the non-random character of the sample of sites.



**Figure 1: Sample of 53 sites in EU-12**

#### **Sampling Sites in Central Europe**

A different approach has been followed to draw a sample of sites in Central Europe (Poland, Czech Republic, Slovakia, Hungary, Romania and Bulgaria) in the frame of the MERA-95 Project (MARS and Environmental Related Applications). The sampling procedure, described in more detail by Vossen et al (1995), has two-phases:

### First sampling phase:

A systematic sample is selected by blocks on the K-J indexes of the "SPOT reference grid". K is related with the satellite orbit, and J refers to the latitude. Sites on the border are selected if more than 50% is inside the studied region. In the example of central Europe, 58 sites were selected in this phase (figure 2). The sampling intensity is higher in the north of the area because the distance  $\eta_{lat(s)}$  between contiguous satellite orbits is shorter.

### Second sampling phase:

Spot XS digital quick-looks or Landsat-TM photographic printouts of the pre-selected sites were analysed for a rough visual estimation  $A_s$  of the percentage of agricultural land. 40 sites were subsampled without replacement with a probability proportional to  $P_s = A_s \times \eta_{lat(s)}$  (figure 3). Estimates are computed with a Horvitz-Thomson or  $\pi$  estimator (Särndal, 1992) using  $1/A_s$  as weights rather than  $1/P_s$  to compensate the variable sampling rate in the first phase.

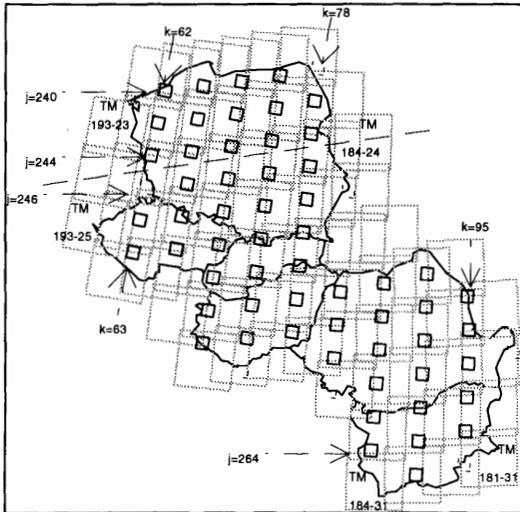


Figure 2: Systematic sample of sites based on the SPOT reference grid.

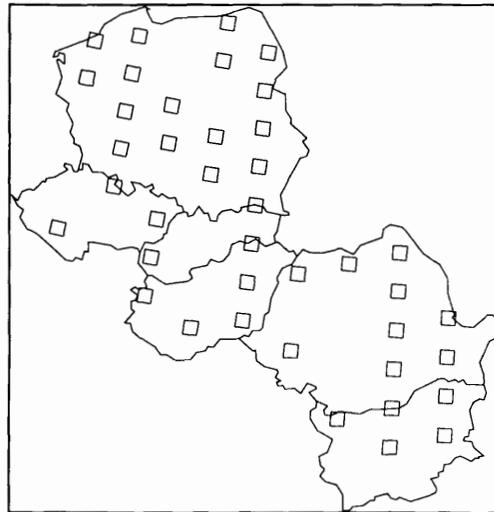


Figure 3: Sample of sites for image analysis

## SAMPLING SEGMENTS IN SATELLITE IMAGE STRIPS

An alternative sampling scheme is currently evaluated. This approach is based on long narrow sites. The site is a strip as long as the studied territory along a satellite orbit track (figure 4). The width of the site can be between 1 km and 10 km, depending on the land use category to be estimated. 1 km or 2 km can suit for agricultural estimates, but wider strips would probably be more efficient in forestry. Different sets of sites can be defined for different satellite sensors, such as Landsat-TM, SPOT-XS, or IRS-1C.

## Sampling Segments inside the Strips

Segment sampling inside the strips follows a two-phase procedure, similar to the site sampling scheme of the previous section. The first phase can be systematic or random; in the simulations presented below we have used systematic sampling, that seems to be generally more efficient in a geographic context (Dunn and Harrison, 1993). In the second phase, a subsample is selected with a probability proportional to an index of agricultural intensity (or forest intensity), and to the distance between satellite orbits.

### Image Acquisition

One important condition for the strip approach to be efficient is that images are marketed by continuous strips instead of approximately square scenes. In the case of Landsat TM, the delivered image for agricultural purpose would have for example a size of  $2000 \text{ km} \times 3 \text{ km} \cong 40 \text{ Mbytes}$  instead of the standard frames  $180 \text{ km} \times 180 \text{ km} \cong 200 \text{ Mbytes}$ .

Strips may be divided into pieces corresponding to different receiving stations: During the agricultural season, image strips would be delivered regardless of the cloud coverage. This gives in Europe approximately 11 TM images or 6-7 SPOT images per year. The selection of images that can be exploited for land use identification in a particular segment would be made later by the photo-interpreter.

The target is getting an “exploitable sample” of at least  $n$  segments. For annual crops, a possible definition of “exploitable segment” can be: *a segment for which at least 3 cloud free images are available during two successive agricultural seasons with a time interval of at least one month.* The definition of exploitable segment is to be adapted to each case; for many forest applications, one cloud free image may be enough. The required number of second phase sample segments to reach the targeted number of exploitable segments is being studied.

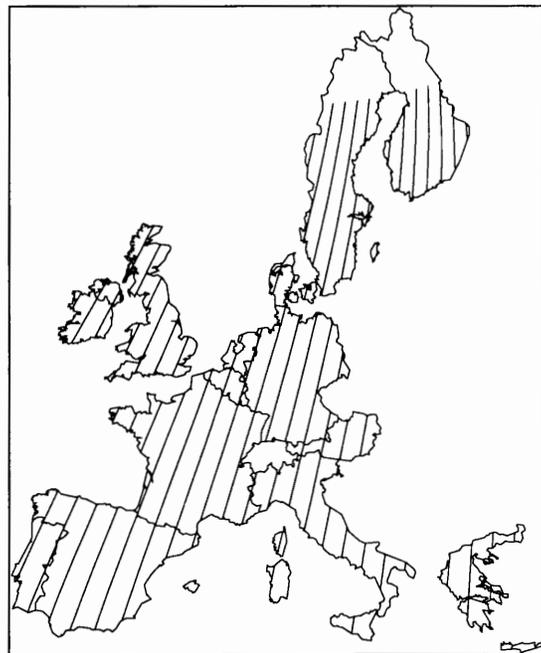


Figure 4: Sites on Landsat tracks

## COMPARING THE EFFICIENCY OF SAMPLING STRATEGIES.

We do not have yet data to test the efficiency of the strip approach with real strips following satellite tracks, but we can simulate the scheme with data from ground survey segments. The set of data used for this test comes from 1992 and 1993 ground surveys on 8023 square segments of 49 ha from the Spanish Ministry of Agriculture in an area of 270.000 Km<sup>2</sup> (Castilla y León, Madrid, Castilla-la Mancha and Andalucía). The set of 8023 segments was obtained by systematic sampling with 3 replicates in blocks of 10 km × 10 km. The sample is described by MAPA (1990), Ambrosio (1993), Fuentes (1994) and Gallego (1995). We consider this set of 8023 segments as population from which we select samples of 80 segments with different strategies (figure 4). Unequal probability subsampling, when applied, was made using 1992 data to compute agricultural intensity indexes. We compare area estimates in 1993 or area change estimates 1993-1992. For each strategy, we have repeated the procedure 100 times in order to estimate variances without unstability problems, that are well known for the variance computation of the Horvitz-Thompson estimator (Cochran, 1977, Särndal, 1992).

### Sampling Strategies Compared

- **Random:** simple random sampling without replacement. This is the reference to compute the relative efficiency of the other strategies as ratio of variances.
- **Systematic:** Systematic sampling ordering population segments by geographic co-ordinates (latitude and then longitude). Uniform probability.
- **Proportional:** Sample with a probability proportional to arable land in 1992.
- **Sites systematic:** Systematic sample by clusters of 16 segments approximately arranged in squares of 40 km × 40 km . Uniform probability.
- **Sites probabilistic:** The same as the previous one with probability proportional to the arable land in 1992. This approach nearly coincides with the one described above, that has been proposed for the MERA-95 Project.
- **Strips:** Two-stage sampling: systematic sampling of strips 100 km apart from each other and sub-sample with probability proportional to 1992 arable land.

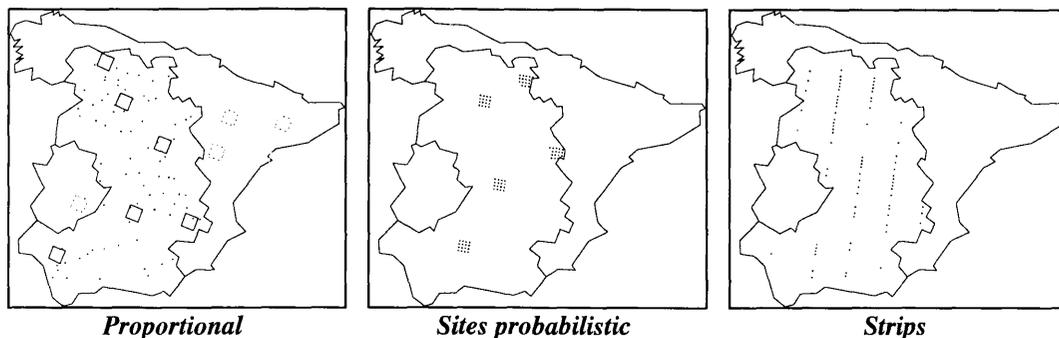


Figure 4: Samples with different strategies

### Comments on Simulation Results.

Some provisional conclusions can be drawn by examining Table 1, that gives the relative efficiency of the studied strategies compared with simple random sampling for the main annual crops and fallow (understood as arable land under repose in the crop rotation).

- Systematic sampling looks slightly superior to random sampling, but there are some important exceptions.
- Sampling unclustered segments with probability proportional to an index of agricultural intensity (strategy *Proportional*) is much more efficient (ratio ranging between 2 and 9) than simple random or systematic sampling. However this strategy may be unsuitable for satellite image acquisition.
- Sampling by square sites of 40 km × 40 km reduces efficiency both with systematic and unequal probability sampling (*Sites systematic* compared with *Systematic* or *Sites proportional* compared with *Proportional*). In the case of systematic sampling, efficiency ratios become close to 1 when area changes are estimated instead of crop area for a particular year. For the more efficient unequal probability sampling, sites of 40 km × 40 km still have a poor behaviour.
- Sampling by strips is not as good as sampling unclustered segments, but is more efficient than sampling by square sites and is compatible with satellite image acquisition.
- The efficiency gain with a two-phase sampling, where the second phase is a subsampling with a probability proportional to arable land is higher than the efficiency obtained in most European countries with area frame sampling stratification (Taylor et al 1996).
- Some exceptions to the comments above can be noted for crops with a relatively small area, as we can check in the case of sugar beet.

**Table 1: relative efficiency compared with random sampling**

	Systematic uniform	Proportional	Sites systematic	Sites probabilistic	Strips
<b>Wheat 93</b>	1.20	6.53	0.36	1.14	2.80
<b>Barley 93</b>	1.61	6.65	0.46	1.12	4.93
<b>Cereals 93</b>	1.42	7.07	0.43	1.47	5.80
<b>Sunflower 93</b>	1.26	3.53	0.20	0.81	1.69
<b>Sugar beet 93</b>	1.14	2.85	0.57	0.66	1.69
<b>Fallow 93</b>	1.05	3.96	0.49	0.49	2.11
<b>Δ Wheat</b>	0.82	2.92	1.36	1.20	1.95
<b>Δ Barley</b>	0.80	2.43	1.01	0.80	2.13
<b>Δ Cereals</b>	1.11	2.39	1.23	0.85	1.94
<b>Δ Sunflower</b>	1.13	2.47	1.13	1.54	2.09
<b>Δ Sugar beet</b>	0.87	3.06	2.42	2.74	1.90
<b>Δ Fallow</b>	1.00	3.09	0.91	0.75	2.27

## CONCLUSIONS

Two-phase sampling may provide a scheme for an adequate use of remote sensing applied to land cover area estimation in large regions, specially if no suitable stratification is available or if available stratification is not very efficient.

Image acquisition by long thin strips instead of approximately square scenes can improve the efficiency of associated sampling plans. However many factors are still to be studied, including image distribution policy, cost of ground surveys and image analysis, and efficiency ratio for different land cover types in different landscapes.

## ACKNOWLEDGEMENTS

We are grateful to the Spanish Ministry of Agriculture, in particular to Porfirio Sánchez, José María Fernández del Pozo, and María José Postigo, that has kindly provided data to test the efficiency of sampling strategies.

## REFERENCES

- Ambrosio L., Alonso R., Villa A., 1993, Estimación de superficies cultivadas por muestreo de áreas y teledetección. Precisión relativa. Estadística Española, Vol. 35, pp. 91-103.
- Carfagna E., Gallego F.J., 1995, Yield Estimates from Area Frame at European Level. Seminar on Yield Forecasting. Villefranche sur Mer, 24-27 October 1994. pp. 237-240.
- Cochran W., 1977, Sampling Techniques. New York: John Wiley and Sons
- Dunn R., Harrison A.R., 1993, Two-dimensional systematic sampling of land use. JRSS: Applied Statistics, vol. 42 n. 4, pp. 585-601.
- Fuentes M., Gallego F.J., 1994, Cluster area estimation on a sample by square grid. getting a stable variance for the survey of the Spanish Ministry of Agriculture. Conference on the MARS Project: overview and perspectives. Office for Publications of the E.C. Luxembourg. pp. 146-149.
- Gallego F.J., Delincé J., Carfagna E., 1994, Two-Stage Area Frame Sampling on Square Segments for Farm Surveys. Survey Methodology , vol 20, No. 2 , pp. 107-115.
- Gallego F.J., 1995, Sampling Frames of Square Segments, Report EUR 16317, Office for Publ. of the E.C. Luxembourg. 68 pp. ISBN 92-827-5106-6

- Kennedy P., Folving S., Mc. Cormick N., 1995, European Forest Ecosystems Mapping and Forest Statistics: the FIRS Project. Proceedings of the workshop on "Defining a System of Nomenclature for European Forest Mapping. June 13-15, 1994, Joensuu, Finland. pp. 33-48. Office for Publications of the EC. Luxembourg.
- Malingreau J.P., Achard F., D'Souza G., Stibig H.T., D'Souza J., Estreguil C., Eva H., 1995, AVHRR for Global Tropical Forest Monitoring: The Lessons of the TREES Project. *Remote Sensing Reviews*, 1995, Vol. 12, pp. 29-40.
- MAPA, 1990, El Marco de Areas como Instrumento de base para la Estadística de superficies de cultivo. *Boletín Mensual de Estad. Min. Agric.* n. 12, pp. 85-106. Madrid
- Mayaux Ph., Lambin E.F., 1995, Estimation of tropical forest area from coarse spatial resolution data: a two-step correction function for proportional errors due to spatial aggregation. *Remote sens. Environ.* n. 53, pp 1-15
- Pous B., Reboux D., Guérif M., Delécolle R., Fischer A., de Boissezon H., 1995. Prévisions du rendement et information satellitaire: acquis de l'action IV du projet MARS et nouvelles perspectives. Seminar on Yield Forecasting. Villefranche sur Mer, 24-27 October 1994. pp. 567-578.
- Särndal C.E., Swenson B., Wretman , 1992, Model Assisted Survey Sampling. Springer Verlag.
- Sharman M., de Boissezon H., (1992), Action IV, de l'image aux statistiques: bilan opérationnel après deux années d' estimation rapides des superficies et des rendements potentiels au niveau Européen. Conference on the Appl. of Remote Sensing to Agricultural Statistics (Belgirate). Office for Publications of the E.C. Luxembourg. pp. 177-186.
- Taylor J., Sannier C., Delincé J, Gallego F.J., 1996, Regional crop inventories in Europe assisted by remote sensing (in press).

## **BIOGRAPHICAL SKETCH**

Javier Gallego graduated in Mathematical Statistics in Valladolid (Spain) and made a PhD. in Multivariate Descriptive Analysis in Paris. He taught Applied Statistics again in Valladolid, where he became Professor in 1986. Then he decided to flee the University and join the European Commission. Since 1988 he works in the Joint Research Centre in Ispra (Italy), where he is in charge of statistical aspects of the MARS Project (Monitoring Agriculture with Remote Sensing).

# Cooperative Accuracy Assessment Strategies for Sampling a Natural Landcover Map of Arkansas

Robert S. Dzur<sup>1</sup>, Michael E. Garner, Kimberly G. Smith, W.  
Fredrick Limp, Donald G. Catanzaro, and Richard L. Thompson

---

**Abstract.**-The National Biological Service's Gap Analysis Program mandates mapping actual vegetation on a state-wide basis. Currently, there are 33 ongoing or completed state gap analysis programs. As those vegetation layers are produced for each state, an increasing concern and need exists to supply some measure of map accuracy. However, challenges associated with designing and implementing systematic validation of state-wide natural landcover maps are significant and potentially cost prohibitive.

A "Rapid Assessment Track" (RAT) for systematic accuracy assessment of a Landsat Thematic Mapper (TM) based natural landcover map of Arkansas was developed and implemented. Arkansas gap analysis adopted a stratified random sampling strategy designed to minimize field data collection costs. Comprehensive mapping goals of the Arkansas gap analysis project warranted development and strong reliance upon state partnership networks. Arkansas Forestry Commission (AFC) personnel conducted actual ground data collection of the systematic sampling network. AFC provided an established and capable field apparatus with state-wide management authority to collect required validation data in a timely manner at reduced cost. Matching the comprehensive Arkansas gap analysis mapping objectives with an equally comprehensive state-wide accuracy assessment would not have been feasible without the cooperative partnership between AFC and Arkansas gap analysis.

---

## INTRODUCTION

Gap analysis combines state-wide Landsat TM derived natural landcover data with predicted terrestrial vertebrate distributions to model biological diversity at eco-regional scales (Scott et al. 1993). Under this scenario, it is important to evaluate accuracy of thematic information, upon which models and potential management decisions may be based.

Error matrices are one commonly used method for assessment of landcover

---

1. Address correspondence to: CAST, University of Arkansas, Ozark Hall Room 12, Fayetteville, AR 72701 (rob@cast.uark.edu)

maps produced by analysis of remotely-sensed digital multispectral satellite data (Congalton 1988, Lark 1995, Verbyla and Hammond 1995). The error matrix is a two-way contingency table that compares classified pixels to a set of independently collected reference data (Janssen and van der Wel 1994).

Collecting unbiased independent reference samples across a 13,798,200 ha study area such as Arkansas becomes problematic. Because of their size, state-wide natural landcover maps offer many difficult logistical issues for systematic thematic accuracy assessments. Among these issues are limited access to private lands, time, and cost constraints. These broad issues characterize the problem of implementing an accuracy assessment on a state-wide scale. While those issues are readily apparent and relatively easy to isolate, the challenge remains to devise available options that address those broad constraints. The Arkansas gap analysis project required an implementation plan - a "Rapid Assessment Track" (RAT) - for gap analysis landcover map accuracy assessment that minimized direct data collection time, costs, and land access limitations.

Perhaps the most influential factor on the magnitude of logistical obstacles lies in the character of the sample design used for accuracy assessment. Janssen and van der Wel (1994) suggested that poor accessibility and limited budgets may result in a cluster-based sampling approach. While clustered sampling might reduce the cost to travel between sample units, field crews still needed to visit and record sample site information. Field crews for accuracy assessment were solicited from Arkansas gap analysis cooperators gathered at the third annual state-wide gap analysis meeting in February of 1995 in Little Rock (Dzur et al. In press). Here, Arkansas Forestry Commission (AFC) demonstrated a willingness to participate in the RAT initiative.

As the state agency responsible for the protection and development of Arkansas' forest resources, AFC possessed key qualities needed to implement RAT. AFC provided a well-developed organizational structure. Furthermore, AFC provided a knowledgeable and well-equipped (vehicles, etc.) personnel base trained in land navigation and natural landcover recognition. This state-wide personnel network facilitated simultaneous sample data collection across the entire state, thus minimizing collection time.

This paper presents development of cooperative implementation strategies for systematic accuracy assessment of a natural landcover map of Arkansas. Research objectives described include: automated sample design development and automated field material production directed toward realization of a simple, systematic, and cost effective state-wide accuracy assessment approach that relied upon an established partnership between the Center for Advanced Spatial Technologies (CAST) and AFC.

## **METHODS**

RAT operational viability resulted from a carefully devised methodology related to sample design and field material composition.

## Sample Site Selection

A critical RAT objective was to employ random sampling at clustered locations. This objective framed another compelling reason to involve AFC in the sampling effort. Commanding state-wide presence and authority, AFC field crews gained ready access to lands otherwise inaccessible due to the geography of land ownership. AFC participation helped enable a random site selection design.

### Site Selection Processing

Sample point selection methods were adopted and modified from Utah gap analysis (Edwards et al. 1995). Arkansas gap analysis spectral classification procedures utilized National Resource Conservation Service (NRCS) State Soil Geographic Data Base (STATSGO) for stratification of spectral space. Accordingly, RAT sampling was stratified by major landforms of Arkansas (figure 1). U. S. Geological Survey (USGS) 7.5 minute quadrangles served as the first filtering strata for the clustering of random samples. A total of 100 USGS 7.5 minute quadrangles (figure 2) were randomly selected in proportion to the relative

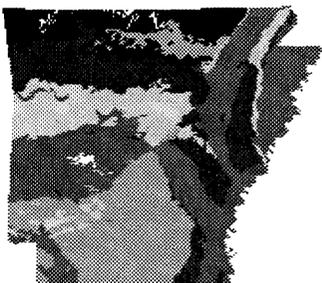


Figure 1. Major Arkansas landforms.

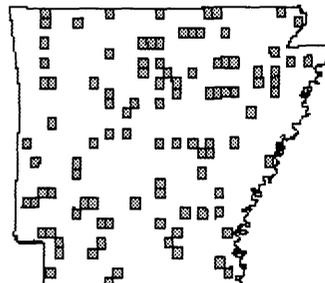


Figure 2. Sample quadrangle locations.

area of Arkansas' 10 major landforms. For example, since the Arkansas River Valley and Ridge landform region comprised 10% of the state, this region was randomly assigned 10 quadrangles.

Once identified, quadrangles were further subdivided into two road-based strata within which a total of 20 1-ha sample points per quadrangle were distributed. Each quadrangle was buffered by 360 m to ensure that roads on surrounding quadrangles did not influence random site selection. Then, the 1:100,000 scale USGS Digital Line Graph (DLG) road network was buffered by 300 m. Within this buffer zone, 10 sites were randomly selected. Next, 10 additional sites were systematically located adjacent to one another based upon a single randomly placed starting point selected outside the roads buffer. These 10 off-road sites were positioned so as not to touch the roads buffer and to lie on a 1 km transect oriented north for easy compass location by the ground crew. A GRASS GIS UNIX script automated random site selection processing for the 100 quadrangles.

### Field Material

While AFC provided the labor force required to visit those sample sites, CAST needed to ensure a quality data collection effort. Accurately locating sample

points in the field was the central concern and most difficult expectation. Since it was impractical to equip AFC with Global Positioning System (GPS) equipment, an alternative was to provide AFC crews with clearly delineated and detailed field map sheets. Designing the most effective field map for AFC personnel required intimate understanding of AFC land navigation training guidelines. AFC outlined these guidelines and recommended a ranked priority of routine AFC land reference features for inclusion on field maps. Since the chain (20 m) is the fundamental unit of field measurement used by AFC personnel, AFC requested a chain map scale. In keeping with AFC needs, maps were printed with a mile barscale subdivided by tenths (80 chains = 1 mile). According to AFC, Public Land Survey System (PLSS) line work was an absolute requirement. Once PLSS data became available from USGS, CAST manufactured a field data collection package complete with detailed field maps, map classification keys, and disposable cameras.

### **Field Map Production**

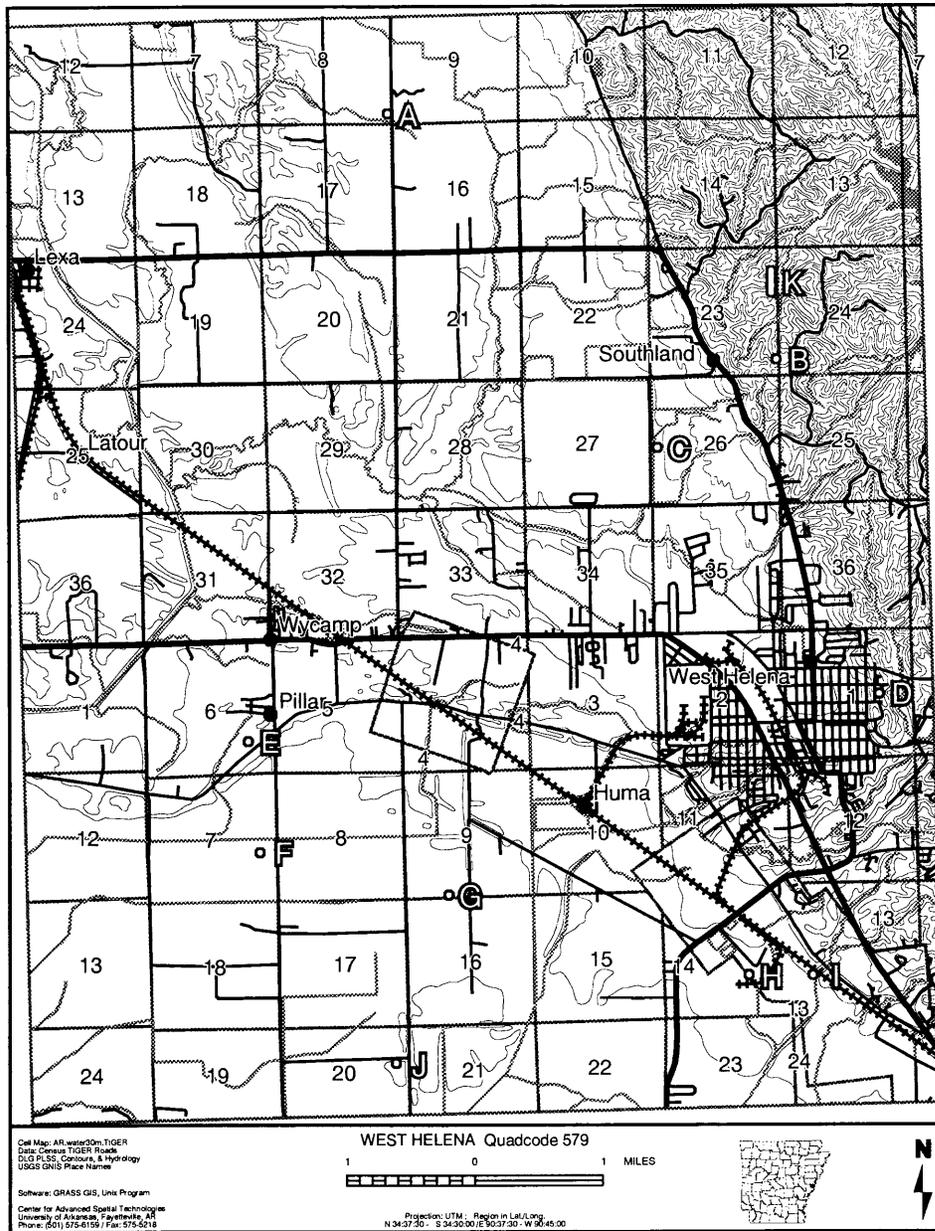
Two sets of 8.5 x 11 field maps (one color and one black and white) were automatically generated for each quadrangle using the Interactive Mapper ([www.cast.uark.edu/products/MAPPER/](http://www.cast.uark.edu/products/MAPPER/)) GRASS script modified to print field map specific features. Distinct color, weight, and style symbology were used to maximize contrast between map features. Separate background raster maps were rendered for color and black and white maps. A Landsat TM unsupervised classification with tasseled-cap (Crist and Cicone 1984) pseudo color table (RGB equals wetness, greenness, brightness) was used as the background raster for color maps. Figure 3, for example, depicts the landscapes of the St. Francis National Forest (upper right) on Crowley's Ridge and the Mississippi alluvial valley. Topologically Integrated Geographic Encoding and Referencing System (TIGER) water was used as the background raster layer for black and white maps (figure 4).

Numerous 1:100,000 scale reference layers were then printed over the raster background. These reference layers appeared on both map sets and included: TIGER roads (with varying line weights and styles for distinct road classes), TIGER railroads, DLG hydrology, DLG contours, and DLG PLSS data (with labelled sections). USGS Geographic Names Information System (GNIS) populated place names and their site locations were printed and oriented such that the names would always be contained entirely within the map area. A white dot surrounded by a black ring marked the 20 sites obtained from the site selection procedure. Sites were given an alphabetical (A-K) letter designation from north to south where K always indicated the 1 km transect. Margin information contained barscale, state locator map, north arrow, quadrangle name and unique number identifier, regional coverage (Latitude/Longitude), source data, and CAST contact information.

A total of 300 (200 color, 100 black and white) maps were delivered to AFC. One set of color maps was maintained at AFC headquarters in Little Rock, AR. The remaining maps were bundled by AFC district for distribution to district foresters and then to actual field crews. To avoid potential bias in data collection,

Figure 3. Color field map.





**Figure 4. Black and white field map.**

no explanation or meaning was attributed to the pattern of color portrayed by the Landsat TM background layer. The black and white map doubled as field data sheet. On the reverse of this map were 20 data entry rows beginning with sample site A. Data entry columns included: date, landcover code, photo #, source, and comments. While on-site visits were preferred, it was recognized that some sites

may have been previously inventoried. Therefore, the source column listed three data survey alternatives: on-site visit (V), inventory records (I), or personal communication (P). At the top of the field data sheet, name and phone number of the field observer were recorded. Upon RAT completion, black and white map/field data sheets were returned to CAST.

### Classification Key Development

A UNESCO/TNC based vegetation classification scheme was adopted for Arkansas gap analysis landcover mapping (Foti et al. 1994). The UNESCO/TNC classification system is characterized by its hierarchical structure (Jennings 1993). To help field crews match ground data to this classification system a coded semi-dichotomous key was assembled from the classification system (figure 5). The key was designed to lead the user through a possible 35 main entries with multiple contrasting sub-entries defining particular landcover criteria to identify the correct landcover map class. Flowing linearly in accordance with the hierarchical structure of the UNESCO based classification, the key first delineates class membership at a system level (e.g., Terrestrial, Developed Cover, Palustrine, Riverine, etc.). The two-page (double-sided) key reads from left to right starting

CODE	DEFINITION	GAP CLASS	GO TO
1(a)	Primarily water (bottomlands)		2
1(b)	Primarily land	TERRESTRIAL	4
1(c)	Primarily altered by humans	URBAN/AGRI/REGEN	35
2(a)	Long term to permanent flooding.		3
2(b)	Seasonal flooding (short to long duration).	PALUSTRINE	23
3(a)	Streamside areas and channels.	RIVERINE	32
3(b)	Wetlands and deep water habitat. 30% tree coverage lacking trees and shrubs. Total area usually > 20 acres.	LACUSTRINE	34
<b>TERRESTRIAL</b>			
4(a)	Areas dominated by trees with a total canopy cover of 61% or more, tree crowns usually interlocking.	FOREST	5
4(b)	Areas dominated by trees with a total canopy cover of 26-60%, most tree crowns not touching each other. An herbaceous or shrub understory, or both are usually present.	WOODLAND	15
4(c)	Areas dominated by shrubs that generally exceed 0.5 meters in height when mature with a total canopy cover of 26% or more. A tree canopy cover of 26% or less may be present.	SHRUBLAND	17
4(d)	Areas dominated by grass, grass-like, or forb vegetation with a tree or shrub component not exceeding 26% cover.	GRASSLAND	18
4(e)	Areas where vegetation cover is less than 5%.	BARE	22
5(a)	Mainly evergreen forest (> than 75% evergreen trees)		6
5(b)	Mainly deciduous or mixed (25% - 75% evergreen trees)		8
6(a)	Dominated by eastern redcedar ( <i>Juniperus virginiana</i> )	T.1.a.9.c.I	
6(b)	Dominated by pines		7
7(a)	Shortleaf pine ( <i>Pinus echinata</i> ) dominant	T.1.A.9.b.I	
7(b)	Loblolly pine ( <i>Pinus taeda</i> ) dominant	T.1.A.9.b.II	
7(c)	Mixed shortleaf pine - loblolly pine (sometimes hickory)	T.1.A.9.b.III	

Figure 5. Example page from Arkansas gap analysis classification key.

with landcover code, followed by a code description, the equivalent gap analysis classification notation (e.g., T.1.A.9.b.I - *Pinus echinata*), and finally a “Go To” column with a number designating the next step down the hierarchy toward a level five (cover type) classification.

To locate shortleaf pine (*Pinus echinata*) dominated forest with the key, for example, AFC field personnel start with code 1a. As 1a - “primarily water (bottomlands)” does not apply, proceeding to 1b - “primarily land” satisfies the present conditions and points to 4. The first entry (4a), “areas dominated by trees with a total canopy cover of 61% or more, tree crowns usually interlocking”, fits the definition of forest leading the search to 5. Again the first entry (5a), “mainly evergreen forest (> than 75% evergreen trees)”, corresponds to ground evidence and leads to 6. As 6a, “dominated by eastern redcedar (*Juniperus virginiana*)”, does not agree, the next alternative, (6b), “dominated by pines”, leads to 7 where 7a, “Shortleaf pine (*Pinus echinata*) dominant”, matches the site characteristics. Then, 7a was recorded on the field data sheet. Instructions to field crews were to record the code about which they had the most confidence.

### Site Verification and Pictorial Key Concept

When landcover codes were listed on the field data sheet, a fundamental question remained. What did *Pinus echinata* (7a) look like? Since data were collected by AFC field crews rather than CAST staff, site characteristics were essentially unknown. To supplement the recorded codes, 25 of the 100 sample quadrangles were arbitrarily allocated a disposable Kodak Fun Camera. At least 2 quadrangles in each major landform unit were assigned cameras. For each sample location, field crews took one picture (figure 6) that best captured the landcover code and recorded the corresponding exposure number in the photo # column. To track camera and quadrangle, a CAST address label printed with quadrangle name and number was attached to each camera. Using RAT as a foundation, these photos might also provide the initial basis for a Web enabled multiscale pictorial key of the Arkansas gap analysis classification that would support written descriptions with pictorial examples of landcover map categories using ground based photos, aerial videography frames, and satellite imagery.

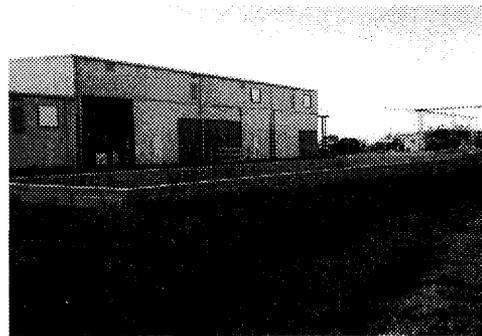


Figure 6. Photo from site H (figure 4).

## RESULTS

A total of 90 field data sheets (quadrangles) were returned. From those quadrangles, AFC collected a total of 1777 field sample points. Two quadrangles lacked the off-road points because they were not printed on the field maps. Only 3

sites were not collected because the landowner would not allow the AFC crew to enter the property. Of the total sample sites, 1601 were actual on-site visits, 156 were based on inventory records, and 20 were based on personal communication.

A total of 23 cameras were returned. From those cameras, 412 photos were developed. Although every roll was not completely developed, most photos appeared very informative. Even underexposed photos provided informational clues related to canopy closure. Expected benefit relative to a total cost of \$330 for camera purchase and film developing justified the experimental camera concept.

## **CONCLUSIONS**

RAT's primary accomplishment was implementing a cooperative state-wide systematic data collection effort. As noted in the methods section, two major elements guided successful RAT implementation. First, the sample design achieved immediate reductions in data collection time through stratification on quadrangles and roads. Second, clear and concise field materials supported AFC field observations. Field materials were generated to assist sample site location, description, and documentation. Furthermore, field materials formed the vital link between sample design and field data collection.

As potential users of the Arkansas natural landcover map, AFC recognized legitimate value in RAT objectives and contributed a proportional commitment to the successful completion of RAT. Although AFC's participation in RAT minimized direct costs to the Arkansas gap analysis project, AFC costs were not minimal. In fact, Robert McFarland, AFC Assistant State Forester, estimated manpower costs alone at one half man year (over \$10,000). Still, RAT demonstrated that partnerships strengthened by well-developed planning can overcome major impediments to state-wide accuracy assessment of land access, collection time, and cost.

## **ACKNOWLEDGMENTS**

The described assessment technique was funded through grants from the United States Fish and Wildlife Service, Award #14-16-0009-1567 RWO #13, as a part of the National Gap Analysis Program, and the National Aeronautics and Space Administration, Award # NAGW-4117 with voluntary contributions from AFC. CAST is appreciative of all AFC field crews that participated in this project. Thanks is due to Robert McFarland (AFC Assistant State Forester) and J. Garner Barnum (AFC Assistant State Forester) for coordinating RAT at AFC and contributing their thorough insight of AFC operations to RAT planning.

## **REFERENCES**

Congalton, C.G. 1988. A comparison of sampling schemes used in generating error matrices for assessing the accuracy of maps generated from remotely sensed data. *Photogrammetric Engineering and Remote Sensing*, 54: 593-600.

- Crist, E.P. and R.C. Cicone. 1984. A physically-based transformation of Thematic Mapper data - The TM Tasseled Cap. *IEEE Transactions on Geoscience and Remote Sensing* GE-22: 256 - 263.
- Dzur, R.S., M.E. Garner, K.G. Smith, and W.F. Limp. In press. Gap Analysis partnerships for mapping the vegetation of Arkansas. In J.M. Scott, T.H. Tear, and F. Davis, eds. *Gap Analysis: A landscape approach to biodiversity planning*. American Society for Photogrammetry and Remote Sensing, Bethesda, Maryland.
- Edwards, T.C. Jr., C.G. Homer, S.D. Bassett, A. Falconer, R.D. Ramsey, and D.W. Wight. 1995. *Utah Gap Analysis: An environmental information system. Final Project Report 95-1*. Utah Cooperative Fish and Wildlife Research Unit. Utah State University, Logan, UT 84322-5210.
- Foti, T., X. Li, M. Blaney, and K.G. Smith. 1994. A classification system for the natural vegetation of Arkansas. *Proceedings Arkansas Academy of Sciences*, 48: 50-62.
- Janssen, L L.F., and F.J.M. van der Wel. 1994. Accuracy assessment of satellite derived land-cover data: A review. *Photogrammetric Engineering and Remote Sensing*, 60: 419-426.
- Jennings, M.D. 1993. Natural terrestrial cover classification: assumptions and definitions. *Gap Analysis Technical Bulletin 2*. U.S. Fish and Wildlife Service, Idaho Cooperative Fish and Wildlife Unit, University of Idaho, Moscow.
- Scott, J.M, F. Davis, B. Csuti, R. Noss, B. Butterfield, C. Groves, H. Anderson, S. Caicco, F. D'Erchia, T.C. Edwards Jr., J. Ulliman, and R.G. Wright. 1993. *Gap Analysis: a geographic approach to protection of biological diversity*. *Wildlife Monographs* 123: 1 - 41.
- Verbyla, D.L., and T.O. Hammond. 1995. Conservative bias in classification accuracy assessment due to pixel-by-pixel comparison of classified images with reference grids. *International Journal of Remote Sensing*, 16: 581-587.

### **BIOGRAPHICAL SKETCH**

Robert S. Dzur is Project Leader for the National Gap Analysis project in Arkansas. He graduated from the University of Arkansas in 1991 with a B.A. double major in Spanish and Geography and received an M.A. in Geography from the University of Arkansas in 1994. His research interests include applications of remote sensing/GIS to scale and spatial dependency of vegetation and landform relationships.

# Fuzzy Measures in Multi-Criteria Evaluation

J. Ronald Eastman<sup>1</sup> and Hong Jiang<sup>2</sup>

---

**Abstract.** -- A consideration of Multi-Criteria Evaluation (MCE) in GIS leads to the conclusion that the standardized factors of MCE belong to a general class of fuzzy measures and the more specific instance of fuzzy sets: a perspective that provides a strong theoretical basis for the standardization of factors and their subsequent aggregation. In this context, a new aggregation operator that permits continuous variation of ANDORness and tradeoff is discussed: the Ordered Weighted Average.

---

## INTRODUCTION

Multi-Criteria Evaluation (MCE) is perhaps the most fundamental of decision support operations in GIS. MCE is most commonly achieved by one of two procedures. The first involves *boolean overlay* whereby all criteria are reduced to logical statements of suitability and then combined by means of one or more logical operators such as intersection or union. The second is known as *weighted linear combination* wherein continuous criteria (*factors*) are standardized to a common numeric range, and then combined by means of a weighted average. The result is a continuous mapping of suitability that may then be masked by one or more boolean *constraints* to accommodate qualitative criteria, and finally thresholded to yield a final decision (Figure 1).

Despite the very common use of these procedures, there are some fundamental problems associated with their use. First, despite a casual expectation that the two procedures should yield similar results, they very often do not (Figure 1). The reason clearly has to do with tradeoff, but also with the logic of the aggregation operation. For example, boolean intersection results in a very hard AND -- a region will be excluded from the result if any single criterion fails to exceed its threshold. Conversely, the boolean union operator implements a very liberal mode of aggregation -- a region will be included in the result even if only a single criterion meets its threshold. Weighted linear combination is quite unlike these options. Here a low score on one criterion can be compensated by a high score on another -- a feature known as *tradeoff* or *substitutability*.

---

<sup>1</sup> Director, The Clark Labs for Cartographic Technology and Geographic Analysis/IDRISI Project, and Professor of Geography, Clark University, Worcester, MA 01610, USA.

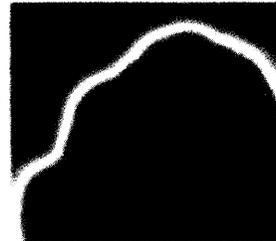
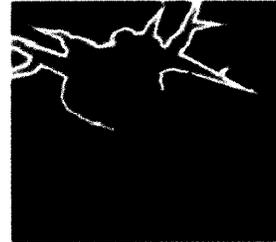
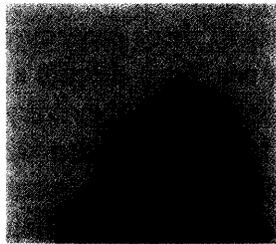
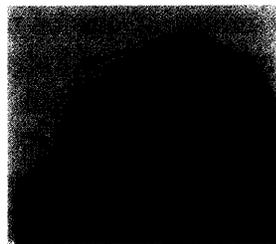
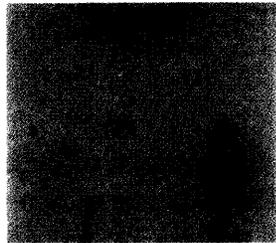
<sup>2</sup> Research Analyst, The IDRISI Project, Clark University, Worcester, MA 01610, USA.

**Figure 1.-GIS procedures for Multi-Criteria Evaluation (MCE)**

**Left: Boolean evaluation for areas suitable for industrial development based on four criteria: proximity to road, proximity to the main town, slope gradient and distance from a wildlife reserve. The bottom is the result for industrial allocation produced by the intersection operator.**

**Right: The same criteria as continuous factors, weighted linear combination, and at the bottom industrial allocation with an area equal to that of the boolean result.**

**Center: At the top is the constraint image with the area that precludes industrial development. On the bottom is the cross-tabulation image of the boolean and weighted linear results.**



The second problem with MCE has to do with the standardization of factors in weighted linear combination. The most common approach is to rescale the range to a common numerical basis by simple linear transformation. However, the rationale for doing so is unclear (Voogd, 1983; Eastman et al., 1993). Indeed, there are many instances where it would seem logical to rescale values within a more limited range. For example, if proximity to roads is a factor in determining suitability for industrial development, it may be that a location 10 km away is no less suitable than one that is 5 km away -- they are both simply too far to consider. Furthermore, there are cases where a non-linear scaling may seem appropriate. Since the recast criteria really express *suitability*, there are many cases where it would seem appropriate that criterion scores asymptotically approach the maximum or minimum suitability level.

A third problem concerns decision risk. Decision risk may be considered as the likelihood that the decision made will be wrong (Eastman, 1996). For both procedures it is a fairly simple matter to propagate measurement error through the decision rule and subsequently to determine the risk that a given location will be assigned to the wrong set (i.e., the set of selected alternatives or the set of those not to be included). However, the continuous criteria of weighted linear combination would appear to express a further uncertainty that is not so easily accommodated. The standardized factors of weighted linear combination each express a perspective of suitability -- the higher the score, the more suitable. However, there is no real threshold that can definitively allocate locations to one of the two sets involved (areas to be chosen and areas to be excluded). How are these uncertainties to be accommodated in expressions of decision risk? If these criteria really express uncertainties, why are they combined through an averaging process? It is the contention here that these and all of the issues raised here can be resolved through a consideration of *fuzzy measures*.

## FUZZY MEASURES

The term *fuzzy measure* refers to any set function which is monotonic with respect to set membership (Dubois and Prade, 1982). Notable examples of fuzzy measures include *probabilities*, the *beliefs* and *plausibilities* of Dempster-Shafer theory, and the *possibilities* of fuzzy sets. Interestingly, if we consider the process of standardization in MCE to be one of transforming criterion scores into set membership statements (i.e., the set of suitable choices), then standardized criteria *are* fuzzy measures.

A common trait of fuzzy measures is that they follow DeMorgan's Law in the construction of the intersection and union operators (Bonissone and Decker, 1986). DeMorgan's Law establishes a triangular relationship between the intersection, union and negation operators such that:

$$T(a, b) = \sim S(\sim a, \sim b)$$

where T = Intersection (AND) = T-Norm  
 and S = Union (OR) = T-CoNorm  
 and ~ = Negation (NOT)

The intersection operators in this context are known as *triangular norms*, or simply *T-Norms*, while the union operators are known as *triangular co-norms*, or *T-CoNorms*.

## T-NORMS/T-CONORMS AND THE AVERAGING OPERATOR

A T-Norm can be defined as (Yager, 1988):

a mapping  $T: [0,1] * [0,1] \rightarrow [0,1]$  such that :

$T(a,b) = T(b,a)$  (commutative)  
 $T(a,b) \geq T(c,d)$  if  $a \geq c$  and  $b \geq d$  (monotonic)  
 $T(a, T(b,c)) = T(T(a,b), c)$  (associative)  
 $T(1,a) = a$

Some examples of T-norms include:

$\min(a,b)$  (the intersection operator of fuzzy sets)  
 $a * b$  (the intersection operator of probabilities)  
 $1 - \min(1, ((1-a)^p + (1-b)^p)^{1/p})$  (for  $p \geq 1$ )  
 $\max(0, a+b-1)$

Conversely, a T-CoNorm is defined as:

a mapping  $S: [0,1] * [0,1] \rightarrow [0,1]$  such that :

$S(a,b) = T(b,a)$  (commutative)  
 $S(a,b) \geq T(c,d)$  if  $a \geq c$  and  $b \geq d$  (monotonic)  
 $S(a, T(b,c)) = T(T(a,b), c)$  (associative)  
 $S(0,a) = a$

Some examples of T-CoNorms include:

$\max(a,b)$  (the union operator of fuzzy sets)  
 $a + b - a*b$  (the union operator of probabilities)  
 $\min(1, (a^p + b^p)^{1/p})$  (for  $p \geq 1$ )  
 $\min(1, a+b)$

Interestingly, while the intersection and union operators of boolean overlay represent a T-Norm/T-CoNorm pair, the averaging operator of weighted linear combination cannot because it lacks the property of associativity (Bonissone and

Decker, 1986). Rather, it has been determined (Bonissone and Decker, 1986) that the averaging operator falls mid-way between the extreme cases of the T-Norm of fuzzy sets (the minimum operator) and its corresponding T-CoNorm (the maximum operator) in the sense that it is neither an AND nor an OR operator, but rather, one that lies halfway in between -- in essence, a perfect ANDOR operator. This quality has interested some in the Decision Science field (e.g., Yager, 1988) because of the recognition that in human perception of decision logics, it is not uncommon to wish to combine criteria with something less extreme than the hard operations of union or intersection. What is particularly interesting here, however, is that the averaging operator falls along a continuum of ANDORness between the intersection and union operators of fuzzy sets. Can the aggregation process of weighted linear combination be seen as fuzzy set membership calculation? Given that it has already been determined that standardized criteria qualify as fuzzy measures, and that fuzzy sets are fuzzy measures, the concept is plausible.

### WEIGHTED LINEAR COMBINATION AND FUZZY SETS

There are a variety of reasons why consideration of the weighted linear combination process as a fuzzy set membership aggregation operator is highly appealing. First, it provides a very strong logic for the process of standardization. In this context, the process of standardizing a criterion can be seen as one of recasting values into a statement of set membership -- the degree of membership in the final decision set. Second, this clearly opens the way for a broader family of set membership functions than that of linear rescaling. For example, the commonly used sigmoidal function provides a simple logic for cases where a function is required that is asymptotic to 0 and 1. Third, the logic of fuzzy sets bridges a major gap between that of boolean overlay and that of the weighted linear combination. Boolean overlay is very clearly a classical set problem. By considering the process as one of fuzzy sets, the process of weighted linear combination can also be seen as a set problem. Fourth, in cases where set membership approaches certainty, the results from fuzzy sets will be identical to those of boolean overlay.

If the problem can indeed be recast into the framework of fuzzy sets, is it reasonable that the minimum and maximum operators should function as the intersection and union operators? As it happens, both are well known in human experience. The minimum operator is commonly used in GIS applications, and represents a form of *limiting factor* analysis. Here the intent is one of risk aversion, by characterizing the suitability of a location in terms of its worst quality. The maximum operator is the opposite, and can thus be thought of as a very optimistic aggregation operator -- an area will be suitable to the extent of its best quality. That the averaging operator should fall half way between these extremes is not surprising.

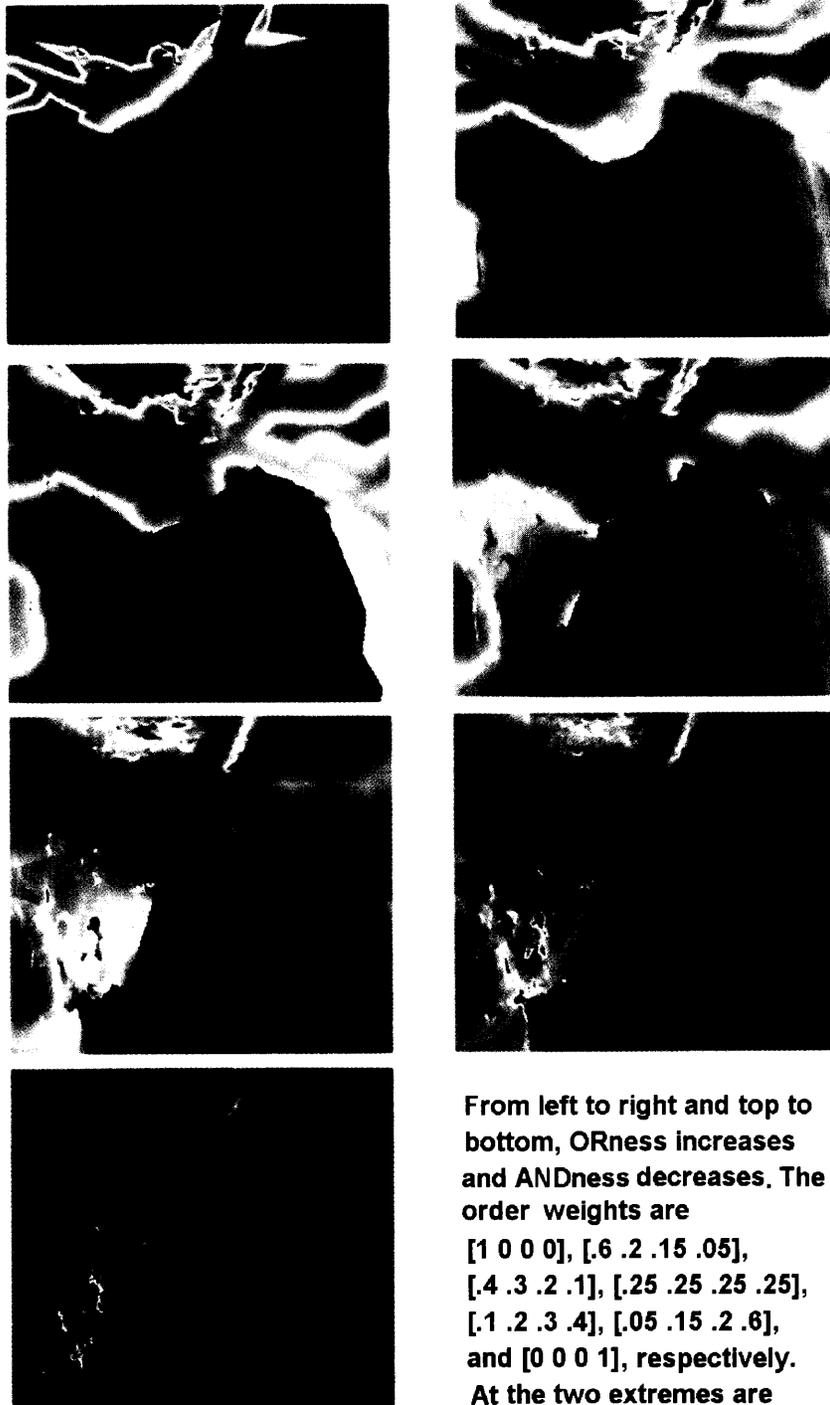


Figure 2.-OWA results with different order weights

## THE ORDERED WEIGHTED AVERAGE

With the recognition that the averaging function represents an ANDOR operator midway between the fuzzy intersection and union operators, and that it also provides full tradeoff between criteria, the question arises of the extent to which these two qualities might be varied independently. Looking to the field of Decision Science, a very interesting approach is that of Yager (1988). Yager has proposed a variant of the averaging operator, called an Ordered Weighted Average, that can achieve continuous control over the degree of ANDORness of the operator and the degree of tradeoff between criteria. In his implementation, the criteria are weighted on the basis of their rank order rather than their inherent qualities. Thus, for example, we might decide to apply weights of 0.5, 0.3, 0.2 to weight a set of factors A, B and C based on their rank order. Thus if at one location the criteria are ranked BAC (from lowest to highest), the weighted combination would be  $0.5*B + 0.3*A + 0.2 *C$ . However, if at another location the factors are ranked CBA, the weighted combination would be  $0.5*C + 0.3*B + 0.2 *A$ . In our implementation of Yager's concept, we have retained the concept of weights that apply to specific factors, yielding two sets of weights -- *criterion weights* that apply to specific criteria and *order weights* that apply to the ranked criteria, after application of the criterion weights.

The interesting feature of the Ordered Weighted Average is that it is possible to control the degree of ANDORness and tradeoff between factors within limits. For example, using order weights of [1 0 0] yields the minimum operator of fuzzy sets, with full ANDness and no tradeoff. Using order weights of [0 0 1] yields the maximum operator of fuzzy sets with full ORness and no tradeoff. Using weights of [0.33 0.33 0.33] yields the traditional averaging operator of MCE with intermediate ANDness and ORness, and full tradeoff. Tradeoff is thus controlled by the degree of dispersion in the weights while ANDness or ORness is governed by the amount of skew. Thus, for example, order weights of [0 1 0] would yield an operator with intermediate ANDness and ORness, but no tradeoff, while the original example with order weights of [0.5 0.3 0.2] would yield an operator with substantial tradeoff and a moderate degree of ANDness.

We have recently implemented the Ordered Weighted Average as a new module (called OWA) to be introduced with the next release of the IDRISI GIS software system. Figure 2 illustrates the output from this module using varying order weights. The results are strikingly different, pointing out the strong dependency between the manner in which the decision rule is constructed and the character of the result.

In use, the OWA module can be used in a single operation, or as part of a hybrid design. For example, if one had five factors, where it was desired that three should trade off, but that two should not, the problem can be solved in two

stages. In the first stage, OWA would be run using order weights with the desired degree of tradeoff, followed by a second run without tradeoff in which the result of the first run was included as an additional factor. Criterion weights in the two runs can easily be adjusted to maintain the original weights of all factors.

## CONCLUSIONS

It has been demonstrated here that the standardized criteria of MCE belong to the general class of fuzzy measures, and it has been further proposed that they belong to the more specific instance of fuzzy sets. The proposal is appealing in that it provides a consistent theoretical link between the MCE logics of boolean overlay and weighted linear combination: the former representing set membership relations between crisp sets and the latter between fuzzy sets. It also provides a theoretical basis for the standardization of criteria: the process is seen as a recasting of criterion values into set membership statements. Furthermore, the simple monotonicity requirement of fuzzy measures permits a much richer range of standardization functions than that of linear rescaling, including sigmoidal and j-shaped forms. However, perhaps the most appealing outcome of this perspective is the manner in which it opens up the possibilities for aggregation of criteria. To this end, the Ordered Weighted Average was demonstrated as an extension to the traditional aggregation operator that allows continuous control over the degree of ANDORness and tradeoff between factors.

## REFERENCES

- Bonissone, P.P. and Decker, K., (1986) "Selecting uncertainty calculi and granularity: an experiment in trading-off precision and complexity. In L.N.Kanal and J.F. Lemmer eds., *Uncertainty in Artificial Intelligence* (North-Holland: Elsevier Science).
- Dubois, D., and Prade, H., (1982) "A Class of Fuzzy Measures based on Triangular Norms", *International Journal of General Systems*, 8, 43-61.
- Eastman, J.R., (1996) "Uncertainty and Decision Risk in Multi-Criteria Evaluation: Implications for GIS Software Design", *Proceedings, UN University International Institute for Software Technology Expert Group Workshop on Software Technology for Agenda'21 : Decision Support Systems*, Feb.26-Mar.8, [in press].
- Eastman, J.R., Kyem, P.A.K., Toledano, J., and Jin, W., (1993) "GIS and Decision Making", *Explorations in Geographic Information System Technology*, 4, (Geneva: UNITAR).
- Voogd, H., (1983) *Multicriteria Evaluation for Urban and Regional Planning*, (London: Pion).
- Yager, R. (1988) "On Ordered Weighted Averaging aggregation operators in multicriteria decision making", *IEEE Transactions on Systems, Man, and Cybernetics*. 8(1): 183-190.

# On Roles and Goals for Map Accuracy Assessment: A Remote Sensing Perspective

Curtis E. Woodcock <sup>1</sup>

---

**Abstract.** – Expanded roles and goals for accuracy assessment are possible based on changes in the assumptions regarding the properties of classes in thematic maps. Existing views of thematic maps use classical set theory, which assumes each location in the map can be assigned unambiguously to a single class. Similarly, classes are assumed exhaustive and mutually exclusive. The substitution of fuzzy set theory for classical set theory relaxes these assumption and affords new opportunities and challenges. The main concept required for use of fuzzy set theory is acknowledgement of varying levels of class membership for individual map units. Three areas where improvements are possible using accuracy assessment based on fuzzy sets are: 1) providing additional information on map accuracy, including information on magnitudes of errors, and alternative methods of estimating the frequency of errors; 2) expanding the kinds of queries that are possible with regard to area estimation, including estimates of areas of map classes as a function of class membership levels; and 3) improved landscape characterization by explicit acknowledgement of the heterogeneity within remote sensing pixels or map polygons. Associated with these expanded roles and goals for accuracy assessment is a need for more research and development.

---

## INTRODUCTION

For a long time thematic maps have been an important method for display of spatial data. Their importance continues, and possibly has been enhanced by the explosive growth in the use of geographic information systems (GIS). There is a clear consensus that accuracy assessment of thematic maps is an important aspect of both map production and use. The importance of accuracy assessment has become particularly apparent due to the use of remote sensing and computer-based methods for production of thematic maps. All maps

---

<sup>1</sup> *Associate Professor of Geography, Boston University Boston, MA.*

contain errors, but the use of more automated methods has brought the issue of map quality and accuracy assessment to the fore.

Accuracy assessment has served two primary communities: the users of maps and the producers of maps. From the production perspective, accuracy assessment is a critical dimension of the process of evaluating alternative mapping methodologies and sensors. Map accuracy assessment has provided the metric for comparisons in a field where comparisons are frequent. From a user's perspective, accuracy assessment provides important information on the quality of the maps, which in turn can significantly influence the manner in which maps can be used. While the motivation for users is typically an indication of the reliability of the map, the dominant concern remains the same -- the overall accuracy of the map.

The purpose of this paper is to argue for expanded roles and goals for accuracy assessment, particularly from the perspective of map production. These more ambitious roles and goals of accuracy assessment are made possible by a change in the fundamental assumptions about the nature of the classes in thematic maps. Conventionally, the classes in thematic maps are assumed to follow classical set theory, in which each location in the landscape is assumed to belong to a single map class. Additionally, the map categories are assumed to be mutually exclusive and exhaustive. Sets, or map classes in this case, that meet these conditions are often termed *crisp sets*. The continuum of variation found in landscapes cannot be captured by classes using the assumptions of crisp sets. The substitution of fuzzy set theory for classical set theory relaxes assumptions regarding the mutually exclusive nature of map classes and affords new opportunities and challenges. I suggest three initial categories of issues arising from the substitution of fuzzy set theory. This list is not intended to be exhaustive, but rather a start. Each of these has received some initial research attention, but more is clearly needed and warranted.

### **Background on Fuzzy Set Theory**

The concept of a fuzzy set was introduced by Zadeh (1965) to describe imprecision that is characteristic of much of human reasoning. Fuzzy sets provide a quantitative approach for dealing with vagueness in complex systems. A primary difference between fuzzy set theory and classical set theory concerns membership functions. In classical set theory, each object or element is either a member of a set or it isn't. Fuzzy set theory allows for grades of membership, providing considerable flexibility beyond that available using classical set theory. For thematic maps, fuzzy sets afford the opportunity to acknowledge ambiguity about the class membership of locations in the map. Sites with properties near the definitional boundaries between classes can be characterized accordingly.

## **ROLES AND GOALS FOR ACCURACY ASSESSMENT**

### **Issue 1: Enhanced Accuracy Information**

Gopal and Woodcock (1994) have provided an initial attempt at accuracy assessment methods for thematic maps based on fuzzy set theory. These methods expand the kind of information resulting from map accuracy assessment. The basic approach to accuracy assessment is similar to conventional methods with respect to sample selection and the need for ground truth (Congalton, 1991). The primary difference in data collection is the requirement for a fuzzy membership value for each possible map class for each sample site. A series of fuzzy operators are used to evaluate the data collected at accuracy assessment sites. Two operators, MAX and RIGHT remain concerned with the primary question of overall map accuracy. However, two independent measures of map accuracy result, with MAX providing a more conservative estimate of map accuracy than RIGHT. The difference in the results for these two operators, combined with the results from a DIFFERENCE operator provide indications of the magnitude of the errors in the map. Conventionally, frequency of errors has been the criterion determining overall map accuracy. Given the ability to measure the magnitude of errors, it would be beneficial to incorporate this information into measures of overall map accuracy.

The nature of the errors in the map, or an analysis of the confusion between classes proceeds in a similar fashion to conventional confusion matrices using a CONFUSION operator. One added dimension associated with the use of fuzzy sets is the AMBIGUITY operator, which tracks the frequency of ties between classes. The results from the AMBIGUITY operator are arranged in a matrix similar to the CONFUSION operator, and are particularly interesting if they are asymmetric. In this situation it is possible to infer the distribution in the map of areas receiving equal fuzzy membership values for two different classes.

Overall, the use of fuzzy operators provides more information about the accuracy of a map than is possible using conventional methods. The use of both the MAX and RIGHT operators increases the information about the frequency of errors, and the overall accuracy of the map. Explicit consideration of the magnitude of errors is also an important development. However, the work of Gopal and Woodcock should be considered only a start. The relaxation of the limitations required by classical set theory provides many opportunities for improved accuracy assessment which remain to be explored.

### **Issue 2: An Expanded View of Area Estimation**

The use of fuzzy sets and the concept of levels of class membership expand significantly the kinds of queries possible regarding the areal extent of classes. Data from conventional map accuracy assessments and marginal map totals have frequently been used to provide improved estimates of the true area of map classes (Card, 1982; Hay, 1988; Jupp, 1989). However, these approaches are limited to the constraint imposed by classical set theory that the sums of the areas of all map classes must be unity. Once the idea that various

grades of class membership exist, it is possible to pose questions regarding the areal extent of classes as a function of level of class membership. When viewed from this perspective, the sums of the areas for all classes need not sum to unity. In fact, values below unity would be expected for strict levels of class membership, and values above unity for more lenient levels of class membership.

Card's (1982) methods for area estimation have been adapted for use in estimating the area in maps having various levels of class membership. The basis for Card's approach is to assign the area in the map represented by each sample (which is a function of the number of samples in the map class, and the map area of the map class) to the true class of the sample site. By retaining the column totals (map classes) for sample sites from the confusion matrix and the marginal map proportions and substituting a confusion matrix that includes only sample sites exhibiting specified levels of class membership, it is possible to estimate the true area of each class for specific levels of class membership (Woodcock and Gopal, 1997). The idea that the sums of the areas of classes might not equal unity is initially unsettling. However, when viewed from outside the normal questions of area estimation, it is easier to understand. For example, one way to view the problem is more like the general question of which locations in a map meet a certain set of conditions. This kind of problem is now common in GIS, and expectations for the results of several such queries to sum to unity rarely exist.

The ability to estimate the areas of classes as a function of levels of class membership should improve the utility of maps. An example from vegetation mapping helps illustrate this point. Currently, users are limited to questions regarding the areas dominated by a particular vegetation type. However, mixes of vegetation types are common. Using the methods mentioned above, it would now be possible to ask questions regarding the areal extent of locations where a particular vegetation type is at least present, if not dominant. These kinds of queries are particularly important in wildlife habitat studies, where the presence of a particular species of tree may be important for nesting purposes.

This direction of research is obviously new, and lacks the benefits of peer review. However, the possibility for improving the information that can be provided to map users is considerable.

### **Issue 3: Improved Landscape Characterization**

One trend in remote sensing has been toward providing maps at regional, continental, and global scales. To accommodate mapping areas of these sizes, increasing use is being made of satellite imagery with coarse spatial resolutions, such as provided by the NOAA AVHRR (see for example, DeFries and Townshend, 1994). One issue arising in the use of coarse spatial resolution imagery in making thematic maps concerns the internal homogeneity of individual pixels. Portraying large areas as a single class provides an imprecise view of landscapes. Prior research has indicated significant errors in the estimation of map proportions can result from the use of large pixel sizes, including the overestimation of the area of larger classes and the disappearance of smaller

classes (Turner et al., 1989; Moody and Woodcock, 1994).

One way to view this problem concerns the ability of classes based on crisp sets to characterize landscapes at these scales. The use of fuzzy sets allows for explicit acknowledgement of internal heterogeneity for large pixels. One community within remote sensing assumes all pixels are mixtures, and employs methods designed to estimate the proportions of constituent components (Adams et al., 1986; Roberts, et al., 1993; Defries et al., 1995). The use of fuzzy set representations for classes in thematic maps may provide a bridge between this mixture modeling community and the production of thematic maps. The key element that fuzzy sets can provide in this context is a method of translating various component mixtures into levels of class membership.

### DISCUSSION AND CONCLUSIONS

This paper calls for a broader role for map accuracy assessment based on the more realistic underlying assumptions of fuzzy set theory. The expanded opportunities carry the requirement of an extensive supporting research agenda. For example, adaptation of some of the recent developments for conventional accuracy assessment for fuzzy sets would be beneficial (Foody, 1992; Ma and Redmond, 1995). While a substantial research agenda is needed, it should be noted that the general steps required for accuracy assessment based on fuzzy sets are quite similar to conventional methods.

The goals of accuracy assessment can be expanded beyond consideration of the nature and frequency of errors in maps to include assessment of the magnitude of those errors and alternative measures of their frequency. Estimation of the area of map classes as a function of class membership levels significantly expands the kinds of queries that are possible, and hence improves the utility of maps. Additionally, the use of fuzzy sets in thematic maps can improve landscape characterization by allowing explicit acknowledgement of heterogeneity within map units.

### REFERENCES

- Adams, J. B., Smith, M.O., and P.E. Johnson, 1986. Spectral Mixture Modeling: A New Analysis of Rock and Soil Types at the Viking Lander 1 Site, *Journal of Geophysical Research*, 91(B8):8098-8112.
- Card, D.H., 1982. Using known map category marginal frequencies to improve estimates of thematic map accuracy, *Photogrammetric Engineering and Remote Sensing*, 48(3):431-439.
- Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data, *Remote Sensing of Environment*, 37:35-46.
- Defries, R.S., and J.R.G. Townshend, 1994. NDVI-derived land cover classifications at a global scale, *International Journal of Remote Sensing*, 15(17):3567-3586.

- DeFries, R.S., Field, C.B., Fung, I., Justice, C.O., Los, S., Matson, P.A., Matthews, P.A., Mooney, H.A., Potter, C.S., Prentice, K., Sellers, P.J., Townshend, J.R.G., Tucker, C.J., Ustin, S.L., and P.M. Vitousek, 1995. Mapping the land surface for global atmosphere-biosphere models: Toward continuous distributions of vegetation's functional properties, *Journal of Geophysical Research*, 100(D10):20,867-20,882.
- Foody, G.M., 1992. On the compensation for chance agreement in image classification accuracy assessment, *Photogrammetric Engineering and Remote Sensing*, 58(10):1459-1460.
- Gopal, S., and C.E. Woodcock, 1994. Theory and Methods for Accuracy Assessment of Thematic Maps Using Fuzzy Sets, *Photogrammetric Engineering and Remote Sensing*, 60(2):181-188.
- Gopal, S., C.E. Woodcock, and G. Unis, 199?. Labeling Map Polygons Using Remote Sensing: Calibrating Decision Rules, submitted to *International Journal of Geographical Information: Analysis, Theory and Decision*, November, 1994, revised August, 1995, revised again November 1995, in press.
- Hay, A.M., 1988. The derivation of global estimates from a confusion matrix, *International Journal of Remote Sensing*, 9:1395-1398.
- Jupp, D.L.B., 1989. The stability of global estimates from confusion matrices, *International Journal of Remote Sensing*, 10(9):1563-1569.
- Ma, Z. and R.L. Redmond, 1995. Tau coefficients for accuracy assessment of classification of remote sensing data, 1995, *Photogrammetric Engineering and Remote Sensing*, 61(4):435-439.
- Moody, A., and C.E. Woodcock, 1994. Scale-Dependent Errors in the Estimation of Land-Cover Proportions -- Implications for Global Land-Cover Datasets, *Photogrammetric Engineering and Remote Sensing*, 60(5):585-594.
- Moody, A., Gopal, S. and Strahler, A., 1995. Sensitivity of neural network outputs to subpixel land-cover mixtures in coarse resolution satellite data, submitted to *Remote Sensing of Environment*, July 1994, revised, September, 1995.
- Roberts, D. A., M. O. Smith, and J. B. Adams, 1993. Green Vegetation, Nonphotosynthetic Vegetation, and Soils in AVIRIS Data, *Remote Sensing of Environment*, 44:222-269.
- Turner, M., O'Neill, R., Gardner, R. and B. Milne, 1989. Effects of changing spatial scale on the analysis of landscape pattern, *Landscape Ecology*, 3(3/4):153-162.
- Woodcock, C.E., and S. Gopal, 199?. Fuzzy Set Theory and Thematic Maps: Accuracy Assessment and Area Estimation, in preparation for *International Journal of GIS*.
- Zadeh, L., 1965. Fuzzy Sets, *Information and Control*, 8:338-353.

# Integration Of Inventory And Field Data For Automated Fuzzy Accuracy Assessment Of Large Scale Remote-Sensing Derived Vegetation Maps In Region 5 National Forests

Jeff A. Milliken<sup>1</sup> and Curtis E. Woodcock<sup>2</sup>

---

**Abstract.**---The U.S.D.A. Forest Service Remote Sensing Lab located in Sacramento, CA is currently developing methods to automate procedures for determining the accuracy of forest-wide regional vegetation maps derived from Landsat Thematic Mapper imagery. Procedures are being developed to use both GPS-controlled FIA (Forest Inventory Analysis) data and additional field data to assess the accuracy of these maps. Methods based on fuzzy set theory are being used in conjunction with this data to determine the nature, magnitude, frequency, and source of error in the maps. This will assess the error for lifeform vegetation labels, CALVEG labels, tree size, and tree cover. The key step under development is the definition of rules for translating field measurements of vegetation properties into fuzzy membership values for the various map classes. GIS layers are utilized for developing necessary data tables for the accuracy assessment procedure.

---

## INTRODUCTION

This paper is a preliminary report of methods currently being developed to define and automate procedures for determining the accuracy of forest-wide regional vegetation maps derived from Landsat Thematic Mapper imagery and GIS modeling.

The Remote Sensing Lab in Region 5 is mapping existing vegetation for the National Forests utilizing a mapping system developed in conjunction with Boston University (Woodcock et al. 1994 and Miller et al. 1994). These vegetation layers are used in the management of the National Forests. For the proper utilization of

---

<sup>1</sup> Remote Sensing/GIS Specialist, U.S. Forest Service Region 5 Remote Sensing Lab, Sacramento, CA

<sup>2</sup> Associate Professor of Geography, Boston University

this data, it is important to provide users with as much information as possible about any error or limitations in the data. All maps have error but it is often difficult to obtain accuracy information on existing data as conducting an accuracy assessment on a data layer can be a cost prohibitive process.

The Remote Sensing Lab is currently developing a methodology with Boston University to assess systematically the accuracy of these vegetation layers using fuzzy sets (Woodcock and Gopal 1992) and an independent field data sample from the Forest Inventory and Analysis (FIA) Permanent Plot Grid System. Utilization of the FIA Grid data will provide for a cost effective means of conducting accuracy assessment.

## **ACCURACY ASSESSMENT METHODS**

### **Field Data**

The FIA Permanent Plot Grid consists of five-point cluster plots sampled on a 3.4 mile grid (U.S. Forest Service 1995). Densification of the grid is sometimes necessary to adequately represent under-sampled forest types. In addition to collecting the cluster plot data, inventory crews also fill out an accuracy assessment form for each point in the plot. This form is intended to capture “expert evaluation” of “best and second best” primary lifeform, CALVEG type (U.S. Forest Service 1981), tree size, and tree density; as well as the same ratings for secondary vegetation types where appropriate. This form may be modified as results from the accuracy assessment methodology are evaluated.

### **Development Of Accuracy Assessment Fuzzy Ratings**

#### **Lifeform and CALVEG**

Map labels from the vegetation layer to be evaluated are added to the field data files by “overlying” the locations of the five-point cluster plots on the vegetation layer in a Geographic Information System. Matrices are developed to compare the map label assigned to each cluster point with the expert evaluations on the accuracy assessment form. Fuzzy ratings are assigned to these unique combinations based on descriptions of the mapped vegetation types, “best and second best” expert evaluations from the accuracy assessment form, and the density of the forest stand where appropriate. For example, in a 20% forest stand with a shrub under-study, a shrub label would get a higher rating of accuracy than in the case of a 60% (denser) forest stand with a shrub label. Fuzzy ratings consist of:

- 5) Absolutely right: No doubt about the match. Perfect.
- 4) Good Answer: Would be happy to find this answer given on the map.
- 3) Reasonable or acceptable answer: Maybe not the best possible answer but it is acceptable.

- 2) Understandable but wrong: Not a good answer. There is something about the site that makes the answer understandable but there is clearly a better answer.
- 1) Absolutely wrong: This answer is absolutely unacceptable.  
(Woodcock and Gopal 1992).

### **Tree size and density**

The Forest Inventory Analysis (FIA) software system allows for the generation of a summary report by position for each cluster plot. This report gives calculated values of mean tree dbh (diameter at breast height), mean crown diameter, and mean canopy cover percent for predominant, dominant and codominant trees in the top-story, and intermediate and suppressed trees in the under-story (U.S. Forest Service 1995). Using these calculated figures may prove to be a more unbiased procedure for assessing the accuracy of tree size and density map labels. Both the mean and their standard errors are used in assigning fuzzy membership values.

### **Final Accuracy Results**

Special considerations will be necessary in reporting final accuracy figures for the nature, magnitude, frequency, and source of error in the maps. Results will need to be weighted with respect to such variables as grid sampling density, representation of vegetation types sampled by the grid, and the five-point cluster plot.

### **CONCLUSIONS**

The accuracy assessment methodology currently under development represents a potentially cost effective means of systematically providing important accuracy information for large scale remote-sensing derived vegetation maps being created in the Region 5 Remote Sensing Lab. This type of information is integral to a user's ability to responsibly utilize these maps for management decisions in the National Forests. The key to the success of this effort is the ability to translate field measurements into fuzzy membership values for map classes.

### **ACKNOWLEDGMENTS**

The authors wish to thank Kevin Casey, John Collins, Kama Kennedy, Ralph Warbington, and Ann Withers for project assistance.

### **REFERENCES**

Miller, S., H. Eng, M. Byrne, J. Milliken, M. Rosenberg. 1994. Northeastern California Vegetation Mapping: A Joint Agency Effort. Proceedings of the Fifth

- Forest Service Remote Sensing Applications Conference. Portland, Oregon. pp. 115-125.
- U.S. Forest Service Regional Ecology Group. 1981. CALVEG: A Classification of California Vegetation. U.S. Department of Agriculture, Forest Service, Region Five. San Francisco, CA. 168 p.
- U.S. Forest Service - Region 5. 1995. Forest Inventory and Analysis User's Guide. U.S. Department of Agriculture, Forest Service, Region Five. San Francisco, CA.
- Woodcock, C.E. and S. Gopal. 1992. Accuracy Assessment of Stanislaus Forest Vegetation Map Using Fuzzy Sets. Proceedings of the Fourth Biennial Remote Sensing Applications Conference. Orlando, Florida. pp. 378-394
- Woodcock, C.E., J.B. Collins, S. Gopal, V. D. Jakabhazy, X. Li, S. Macomber, S. Ryherd, Y. Wu, V.J. Harward, J. Levitan, and R. Warbington. 1994. Mapping Forest Vegetation Using Landsat TM Imagery and a Canopy Reflectance Model. Remote Sensing of Environment, 50:240-254.

### **BIOGRAPHICAL SKETCH**

Jeff A. Milliken is a geologist and remote sensing specialist in the U.S. Forest Service Remote Sensing Lab located in Sacramento, CA. He graduated from Colorado State University in 1975 with a B.S. in Geology and is currently completing his graduate work in Geography/Remote Sensing at San Francisco State University.

Curtis E. Woodcock is an Associate Professor of Geography at Boston University. His academic degrees are from the University of California, Santa Barbara, with interests in many aspects of remote sensing.

# Uncertainty of Spatial Metric Relations in GIS

Xiaoyong Chen<sup>1</sup>, Takeshi Doihara<sup>1</sup> and Mitsuru Nasu<sup>1</sup>

---

**Abstract.**---In this paper, after an introduction to the notations of *metric topology*, a novel theory of spatial metric relations between sets is developed in which the relations are defined in terms of the intersections of the boundaries, interiors and exteriors of two dynamically generated sets. Then, the presented theory is extended to quantitatively deriving the spatial metric relations between sets in consideration of conceptual and positional uncertainties based on the fuzzy set theory. Finally, some potential applications of presented theories and the ideas for spatial and temporal reasoning in GIS are also suggested.

---

## INTRODUCTION

One of the most fundamental requirements for GIS is to modeling and communicating error in spatial databases. With increased research into error modeling in GIS over the past few years, there has a considerable body of models and techniques available for measurement spatial database error from researching to real applications [Goodchild and Gopal, 1989; Vergin, 1994]. Spatial relationships (such as distance, direction, ordering, and topology) between objects, as very useful tools for spatial queries, constraints checking and sorting in GIS, are strongly influenced by the uncertainties of original spatial data. These practical needs in GIS have led to the investigation of formal and sound methods for driving spatial relations and their variations with uncertainties [Chen and et.al., 1995, 1996; Egenhofer and Franzosa, 1991; Frank, 1992; Kainz, et.al., 1993; Mark, et. al., 1995; Peuquet and Zhang, 1987]. However, how to derive spatial relations between sets (non-point-like) in consideration of conceptual and positional uncertainties based on an mathematically well-defined algebra framework is still an open problem up to now. The lack of this comprehensive theory has been a major impediment for solving many sophisticated problems in GIS, such as formally deriving complex spatial relations among spatial objects with multiple representations, and generation of the related standards for transferring spatial relations.

This paper focuses on the development of the theory and associated models of spatial metric relations between sets in consideration of data uncertainties. At first, based on the metric topology, a novel theory of spatial metric relations between sets is developed in which the relations are defined in terms of the intersections of the boundaries, interiors and exteriors of two dynamically generated sets. Then, the

---

<sup>1</sup> AAS Research Institute, Asia Air Survey Co., LTD.,8-10, Tamura-Cho, Atsugi-Shi, Kanagawa 243, JAPAN

presented theory is extended to quantitatively deriving the spatial metric relations between sets in consideration of conceptual and positional uncertainties based on the fuzzy set theory. Finally, some potential applications of presented theories and the ideas for spatial and temporal reasoning in GIS area are also suggested.

## FUNDAMENTAL DEFINITIONS

### Metric Spaces

A *metric space* is a pair consisting of a set  $E$  and a mapping  $(p_1, p_2) \rightarrow d(p_1, p_2)$  of  $E \times E$  into  $R_+$ , having the properties: (1).  $p_1 = p_2 \Leftrightarrow d(p_1, p_2) = 0$ ; (2).  $d(p_1, p_2) = d(p_2, p_1)$  (*symmetry*); (3).  $d(p_1, p_2) \leq d(p_1, p_3) + d(p_3, p_2)$  (*triangle inequality*). The function  $d$  is called a *metric* and  $d(p_1, p_2)$  is called the *distance* between the points  $p_1$  and  $p_2$ . Distance between points  $p_i(x_{i1}, x_{i2}, \dots, x_{in})$  in  $R^n$  is described in terms of the *Minkowski*  $d_l$ -metric:

$$d_l(p_1, p_2) = \left( \sum_{j=1}^n |x_{1j} - x_{2j}|^l \right)^{1/l} \quad [1]$$

Conventional *Euclidean distance* is defined by the  $d_2$ -metric. Similarly, the *Manhattan distance* defined by the  $d_1$ -metric, and the *maximum distance* defined by the  $d_\infty$ -metric.

### Metric Topological Spaces

#### Topological spaces

A *topological space* is a pair consisting of a set  $E$  and a collection  $A$  of subsets of  $E$  called the *open sets*, satisfying the three following properties: (1). every union (finite or otherwise) of open sets is open; (2). every finite intersection of open sets is open; (3). the set  $E$  and the empty set  $\emptyset$  are open.

#### Metric topology

A metric  $d$  on a set  $E$  includes a topology on  $E$ , called *metric topology* defined by  $d$ . This topology is such that  $U \subset E$  is an open set if, for each  $p_1 \in U$ , there is an  $\varepsilon > 0$  such that the  $d$ -ball of radius  $\varepsilon$  around  $p_1$  is contained  $U$ . A  $d$ -ball is the set of points whose distance from  $p_1$  in the metric  $d$  is less than  $\varepsilon$ , i.e.  $\{p_2 \in E | d(p_1, p_2) < \varepsilon\}$ . Notice that the metric topological spaces are *Hausdorff* and *separable*.

### The Hausdorff Metric

The Hausdorff metric is defined on the space  $K$  where each point is a non empty compact set of  $R^2$ . If  $K_1$  and  $K_2$  denote two non empty compact set in  $R^2$  (or equivalently two points in  $K$ ), and  $B(\varepsilon)$  is the closed ball with a radius  $\varepsilon$ , then the quantity:

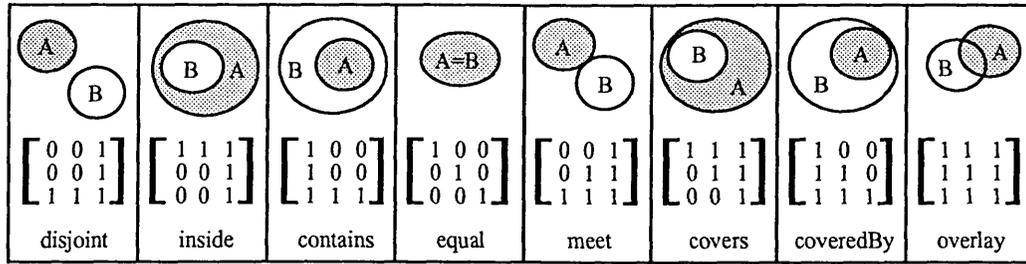


Figure 1. ---The eight topological relations between regions in  $IR^2$  .

$$\rho(K_1, K_2) = \inf\{\varepsilon: K_1 \subset K_2 \oplus B(\varepsilon), K_2 \subset K_1 \oplus B(\varepsilon)\} \quad [2]$$

defines a metric  $\rho$  on  $K$  , known as the Hausdorff metric. From equation [2],  $\rho$  is the radius of the smallest closed ball  $B$  such that both  $K_1$  is contained in the set  $K_2 \oplus B(\varepsilon)$  generated by the morphological dilation [Serra, 1982] and  $K_2$  is contained in the dilated set  $K_1 \oplus B(\varepsilon)$  . Serra (1982) has proven that the Hausdorff distance  $\rho(K_1, K_2)$  satisfies all properties of distance functions. In particular case, when  $K_1$  and  $K_2$  are reduced to two points, the Hausdorff distance  $\rho(K_1, K_2)$  coincides with the Euclidean distance.

## SPATIAL METRIC RELATIONS BETWEEN SETS

### Set Intersection Models

#### 9-Intersection

For driving binary topological relations between sets, Egenhofer et al., (1994) developed the 9-intersection model based on the usual concepts of point-set topology with open and closed sets, in which the binary topological relations between two objects,  $K_1$  and  $K_2$ , in  $IR^2$  is based upon the intersection of  $K_1$ 's interior ( $K_1^\circ$ ), boundary ( $\partial K_1$ ), and exterior ( $K_1^-$ ) with  $K_2$ 's interior ( $K_2^\circ$ ), boundary ( $\partial K_2$ ), and exterior ( $K_2^-$ ). A  $3 \times 3$  matrix  $\mathfrak{I}_9$ , called the 9-intersection as follows:

$$\mathfrak{I}_9 = \begin{bmatrix} K_1^\circ \cap K_2^\circ & K_1^\circ \cap \partial K_2 & K_1^\circ \cap K_2^- \\ \partial K_1 \cap K_2^\circ & \partial K_1 \cap \partial K_2 & \partial K_1 \cap K_2^- \\ K_1^- \cap K_2^\circ & K_1^- \cap \partial K_2 & K_1^- \cap K_2^- \end{bmatrix} \quad [3]$$

By considering the values empty (0) and non-empty (1) in equation [3], one can distinguish between  $2^9=512$  binary topological relations in which only a small subset can be realized when the objects of concern are embedded in  $IR^2$  . Egenhofer and Franzosa (1991) showed that, for two regions with connected boundaries embedded in  $IR^2$  , the 9-intersection distinguishes just 8 different relations, i.e. *disjoint*, *contains*, *inside*, *equal* *meet*, *covers*, *coveredBy*, and *overlap* (figure 1). However, when we apply the 9-intersection model to describing topological relations between other types of spatial objects, the situation will be more complicated. According to the results of Mark et. al. (1995), for two simple lines 33 different spatial relations are

possible, and for a line and a region, 19 are possible.

### Dynamic 9-intersection

For integrally deriving different kinds of spatial relations between sets, Chen et al., (1995, 1996) developed the dynamic 9-intersection model based on the concepts of the metric topology with open and closed sets and the morphological dilation, in which the general 9-intersection of equation [3] is extended as follows:

$$\mathfrak{I}_{(1,2)}^9(\varepsilon_i) = \begin{bmatrix} [K_1 \oplus B(\varepsilon_i)]^\circ \cap K_2^\circ & [K_1 \oplus B(\varepsilon_i)]^\circ \cap \partial K_2 & [K_1 \oplus B(\varepsilon_i)]^\circ \cap K_2^- \\ \partial [K_1 \oplus B(\varepsilon_i)] \cap K_2^\circ & \partial [K_1 \oplus B(\varepsilon_i)] \cap \partial K_2 & \partial [K_1 \oplus B(\varepsilon_i)] \cap K_2^- \\ [K_1 \oplus B(\varepsilon_i)]^- \cap K_2^\circ & [K_1 \oplus B(\varepsilon_i)]^- \cap \partial K_2 & [K_1 \oplus B(\varepsilon_i)]^- \cap K_2^- \end{bmatrix} \quad [4]$$

where the  $K_1$  and  $K_2$  are given two closed sets, the  $K_i \oplus B(\varepsilon_i)$  means relevant morphological dilation by the closed ball  $B$  with radius  $\varepsilon_i$ , and the  $\mathfrak{I}_{(i,j)}^9(\varepsilon_i)$  means dynamic 9-intersection with parameter  $\varepsilon_i$  from  $K_i$  to  $K_j$ . Based on the equation [4], we can derive dynamic topological relations by using the different parameter  $\varepsilon_i$ .

In particular case, when  $\varepsilon_i=0$ , the structure element  $B(\varepsilon_i)$  is reduced to the original point  $\{o\}$ , according to the algebraic properties of morphological dilation [Serra, 1982], we have  $K_i \oplus B(\varepsilon_i) = K_i \oplus \{o\} = K_i$ , then the dynamic 9-intersections  $\mathfrak{I}_{(i,j)}^9(\varepsilon_i)$  defined in equation [4], coincide with the general 9-intersection  $\mathfrak{I}_9$  in equation [3].

## Spatial Metric Relations between Sets

### Distance Relations between Sets

According to the derived dynamic topological relations by the dynamic 9-intersections of equation [4] with different parameter  $\varepsilon_i$ , such as *dynamic equal*, *dynamic covers* (or *dynamic coveredBy*) and *dynamic contains* (or *dynamic inside*), we can easily get the Hausdorff distance  $\rho(K_1, K_2)$  between two closed sets  $K_1$  and  $K_2$  by calculating the minimum and maximum dilated distances:

$$\rho(K_1, K_2) = \max\{\min(\varepsilon_i), \min(\varepsilon_j)\}; \text{ when } \left. \begin{array}{l} \mathfrak{I}_{(1,2)}^9(\varepsilon_i) = \left[ \begin{array}{c} \begin{bmatrix} 100 \\ 010 \\ 001 \end{bmatrix} \begin{bmatrix} 000 \\ 010 \\ 001 \end{bmatrix}, \begin{bmatrix} 111 \\ 001 \\ 001 \end{bmatrix} \begin{bmatrix} 011 \\ 001 \\ 001 \end{bmatrix}, \begin{bmatrix} 001 \\ 011 \\ 001 \end{bmatrix} \begin{bmatrix} 001 \\ 111 \\ 001 \end{bmatrix}, \begin{bmatrix} 001 \\ 101 \\ 001 \end{bmatrix} \begin{bmatrix} 111 \\ 001 \\ 001 \end{bmatrix}, \begin{bmatrix} 1*1 \\ *11 \\ 001 \end{bmatrix} \right] \\ \mathfrak{I}_{(2,1)}^9(\varepsilon_j) = \left[ \begin{array}{c} \begin{bmatrix} 100 \\ 010 \\ 001 \end{bmatrix} \begin{bmatrix} 000 \\ 010 \\ 001 \end{bmatrix}, \begin{bmatrix} 100 \\ 100 \\ 111 \end{bmatrix} \begin{bmatrix} 000 \\ 100 \\ 111 \end{bmatrix}, \begin{bmatrix} 000 \\ 010 \\ 111 \end{bmatrix} \begin{bmatrix} 000 \\ 100 \\ 111 \end{bmatrix}, \begin{bmatrix} 010 \\ 010 \\ 111 \end{bmatrix} \begin{bmatrix} 110 \\ 100 \\ 111 \end{bmatrix}, \begin{bmatrix} 1*0 \\ *10 \\ 111 \end{bmatrix} \right] \end{array} \right\} \quad [5]$$

where "\*" means either empty (0) or non-empty (1). The binary distance relations derived by equation [5] are suitable for different types of spatial objects, such as point-objects, line-objects and region-objects, as well as combining different types of spatial objects such as a line and a region, a point and a line, or a point and a region. Some examples are shown in figure 2.

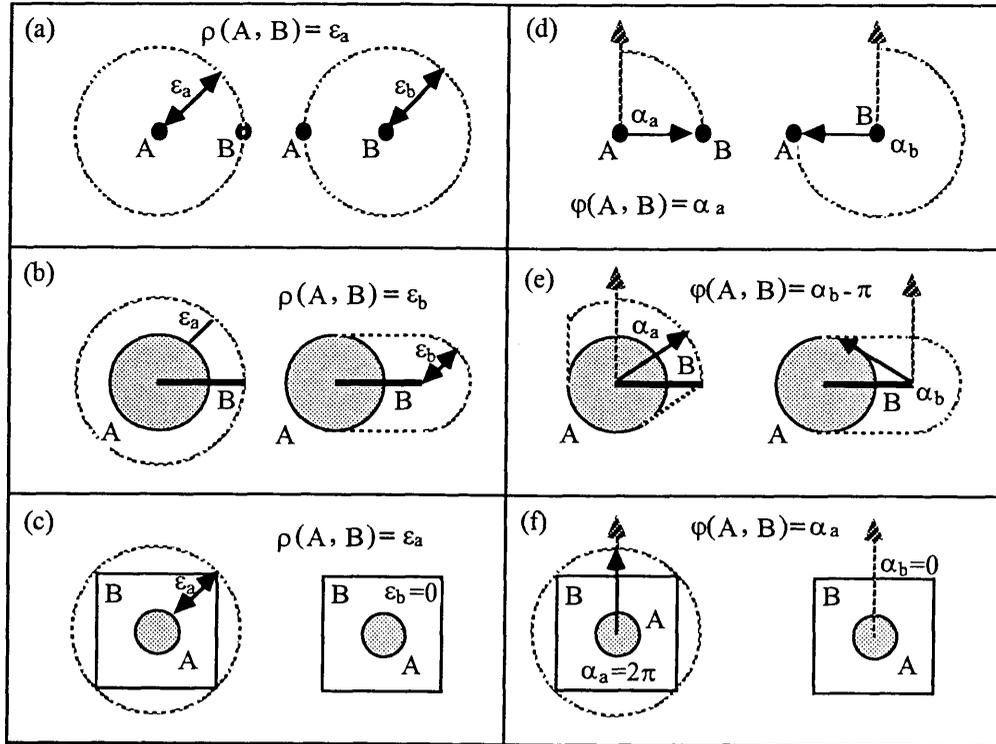


Figure 2. ---The spatial metric relations between two planar sets A and B.

### Directional Relations between Sets

Directional relations between sets can be defined by the Hausdorff metric of *angular bearings*, in which the direction from one spatial object to another is identical to that for metric function except that the angular bearing is computed for each ordered pair in the Cartesian product. The angular bearing is measured in the sense of navigation bearings (i.e. increasing clockwise from north).

Similarly, for calculation of the directional relation  $\varphi(K_i, K_j)$  between two non empty compact sets  $K_1$  and  $K_2$  in  $R^2$ , we can select the angular bearing set  $R(\alpha_i)$  as instead of the closed ball  $B(\varepsilon_i)$  in equation [3], then we can get the Hausdorff direction  $\varphi(K_i, K_j)$  by calculating the minimum and maximum dilated angles as follows:

$$\left. \begin{aligned} \varphi(K_1, K_2) &= \max\{\min(\alpha_i), \min(\alpha_j \pm \pi)\} \\ \varphi(K_2, K_1) &= \varphi(K_1, K_2) \pm \pi \end{aligned} \right\}; \text{ when}$$

$$\left. \begin{aligned} \mathfrak{F}_{(1,2)}^{\circ}(\alpha_i) &= \left[ \begin{array}{c} 100 \\ 010 \\ 001 \end{array} \right] \left[ \begin{array}{c} 000 \\ 010 \\ 001 \end{array} \right], \left[ \begin{array}{c} 111 \\ 001 \\ 001 \end{array} \right] \left[ \begin{array}{c} 011 \\ 001 \\ 001 \end{array} \right], \left[ \begin{array}{c} 001 \\ 011 \\ 001 \end{array} \right], \left[ \begin{array}{c} 001 \\ 111 \\ 101 \end{array} \right] \left[ \begin{array}{c} 111 \\ 101 \\ 001 \end{array} \right] \left[ \begin{array}{c} 1*1 \\ *11 \\ 001 \end{array} \right]; \\ \mathfrak{F}_{(2,1)}^{\circ}(\alpha_j) &= \left[ \begin{array}{c} 100 \\ 010 \\ 001 \end{array} \right] \left[ \begin{array}{c} 000 \\ 010 \\ 001 \end{array} \right], \left[ \begin{array}{c} 100 \\ 100 \\ 111 \end{array} \right] \left[ \begin{array}{c} 000 \\ 100 \\ 111 \end{array} \right] \left[ \begin{array}{c} 000 \\ 010 \\ 111 \end{array} \right], \left[ \begin{array}{c} 010 \\ 010 \\ 111 \end{array} \right] \left[ \begin{array}{c} 110 \\ 100 \\ 111 \end{array} \right] \left[ \begin{array}{c} 1*0 \\ *10 \\ 111 \end{array} \right]. \end{aligned} \right\} \quad [6]$$

equal
contains
covers

where "\*" means either empty (0) or non-empty (1). The binary directional relations derived by [6] are also suitable for different types of spatial objects and their combining types. Some examples are shown in figure 2.

## UNCERTAINTY OF SPATIAL METRIC RELATIONS

### Conceptual Uncertainty of Spatial Metric Relations

Unlike the spatial metric relations between points, the concepts of distances and directions between subsets are fuzzy [Zadeh, 1965], since the spatial objects may contain many subsets, the distances and directions between these subsets are difficult to be represented just by a single value. If we define the fuzzy memberships as the covering percentages of generated region areas ( or point numbers and line lengths ) by the dynamic intersection of sets in  $R^2$ , we can find that the Hausdorff metric is just the special case with the fuzzy membership value equal to one. Based on the changes of the covering areas of regions (or point numbers and line lengths) from an empty set to a complete set, we can estimate the fuzzy memberships from zero to one, then we can quantitatively derive the spatial metric relations between subsets.

For reasons of simplicity the distances and directions between closed sub-regions discussed in this paper only, related models for estimation of conceptual fuzzy membership functions are defined as follows:

$$\left. \begin{aligned} \Phi_{\lambda}(\lambda) &= \frac{A\{[K_1 \oplus B(\lambda)] \cap K_2\} + A\{K_1 \cap [K_2 \oplus B(\lambda)]\}}{A(K_1) + A(K_2)} \\ \Omega_{\theta}(\theta) &= \frac{A\{[K_1 \oplus R(\theta)] \cap K_2\} + A\{K_1 \cap [K_2 \oplus R(\theta \pm \pi)]\}}{A(K_1) + A(K_2)} \end{aligned} \right\} [7]$$

where  $A\{*\}$  means covered area sizes by dynamic intersections with the parameters  $\lambda$  and  $\theta$  for distances and directions separately, the functions  $0 \leq \Phi_{\lambda}(\lambda) \leq 1$  and  $0 \leq \Omega_{\theta}(\theta) \leq 1$  with the parameters  $0 \leq \lambda \leq \rho(K_1, K_2)$  and  $0 \leq \theta \leq \varphi(K_1, K_2) \pm \pi$  are called the size distribution functions [Serra, 1982]. An example of conceptual uncertainty of spatial distance relations between two spatial regions is shown in figure 3.

### Positional Uncertainty of Spatial Metric Relations

Generally, the locations of objects in spatial databases are not error-free, they may contain many kinds of errors, such as the errors of scanning, digitizing, selecting, projection, overlaying, and et.al. [Goodchild and et.al., 1989]. For describing the uncertainty in the positions of spatial objects (such as lines and areas), we can use the error model of  $\varepsilon$ -band developed by Chrisman (1982), in which the positional uncertainty of a spatial object  $K_i$  can be represented as  $\tilde{K}_i = K_i \oplus \mu(\varepsilon)$ , where  $\varepsilon$  is the buffer distance of error distribution and  $\mu(\varepsilon)$  is the fuzzy membership function derived by  $\varepsilon$ . According to the equations of [2]-[7], if we use the uncertain spatial object  $\tilde{K}$  as instead of the general spatial object  $K_i$  in the equations of [2]-[7], we

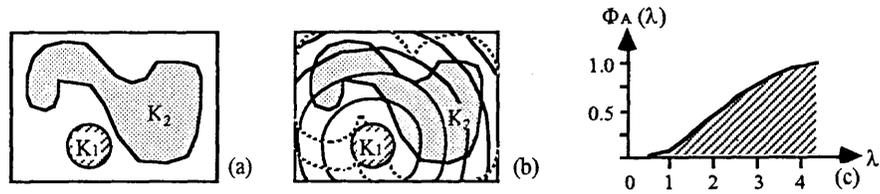


Figure 3. ---The conceptual fuzzy membership function between subsets.

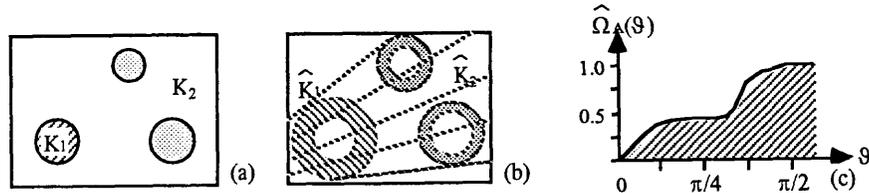


Figure 4. ---The positional fuzzy membership function between subsets.

can drive the spatial metric relations between uncertain sets and the conceptual fuzzy membership functions between uncertain subsets. For quantitatively estimating the positional fuzzy membership functions of derived spatial metric relations between uncertain subsets, we can use the measurement of dynamic covering uncertain areas as follows:

$$\left. \begin{aligned} \hat{\Phi}_A(\lambda) &= \frac{A\{[\tilde{K}_1 \oplus B(\lambda)] \cap \hat{K}_2\} + A\{\hat{K}_1 \cap [\tilde{K}_2 \oplus B(\lambda)]\}}{A(\hat{K}_1) + A(\hat{K}_2)} \\ \hat{\Omega}_A(\theta) &= \frac{A\{[\tilde{K}_1 \oplus R(\theta)] \cap \hat{K}_2\} + A\{\hat{K}_1 \cap [\tilde{K}_2 \oplus R(\theta \pm \pi)]\}}{A(\hat{K}_1) + A(\hat{K}_2)} \end{aligned} \right\} [8]$$

where  $\hat{K}_i = \tilde{K}_i - K_i = [K_i \oplus \mu(\varepsilon)] - K_i$  means the uncertain area generated by the related  $\varepsilon$ -band,  $A\{*\}$  means the covered size of area or volume when  $\mu(*)$  is a constant or a general fuzzy member function with the value  $0 \leq \mu(*) \leq 1$  separately. An example of positional uncertainty of spatial directional relations between two spatial regions is shown in figure 4.

## CONCLUSIONS AND OUTLOOKS

As the natural extension of the general 9-intersection, the dynamic 9-intersection based on metric topology supplied a general framework for studying metric spatial relations between sets. The presented integrated theory of spatial relations between sets makes a new way for formally deriving spatial relations among complex spatial objects with conceptual and positional uncertainties. Even though the presented approach is only focus on the applications in GIS field, the related results for deriving spatial relations between sets and uncertain sets can be also used for many other fields, such as CAD, computer vision, pattern recognition, robot space searching and so on. However, only the theoretical models have be presented in this

paper, a wide field of practical application for management and analysis of spatial data in 2-D and 3-D GIS environments has not been touched. Therefore, the reported results must be verified and extended in order to be used in different practical environments.

Two main directions for further research shall be pointed here, one is the applications of the presented theoretical models and algorithms in 2-D and 3-D GIS environments for developing the new tools of spatial query and analysis; another one is the extensions of presented theories and models for formally deriving complex spatial relations among spatial objects with multiple representations.

### ACKNOWLEDGMENTS

The authors wish especially to thank Prof. S.Murai and Mr. H.Yamamoto for their considerable helps and supports of this research project.

### REFERENCES

- Chen, X., et. al., 1995. Spatial Relations of Distances between Arbitrary Objects in 2-D/3-D Geographic Spaces based on the Hausdorff Metric. LIESMARS'95, Wuhan, China. 30-41 p.
- Chen, X., 1996. Spatial Relations between Sets. ISPRS XVIII Congress. Vienna. Austria. (accepted)
- Chrisman, N. R., 1982. Methods of spatial analysis based on errors in categorical maps. Unpublished PhD thesis, University of Bristol.
- Egenhofer, M., and Franzosa, R., 1991. Point-Set Topological Spatial Relations. IJGIS. 5(2), 161-174 p.
- Egenhofer, M., and et. al., 1994. A Critical Comparison of the 4-Intersection and 9-Intersection Models for Spatial Relations: Formal Analysis, AUTO CARTO 11. 1-11 p.
- Frank, A., 1992. Qualitative Spatial Reasoning about Distances and Directions in Geographic Space. J. of Visual Languages and Computing. 3(4), 343-371 p.
- Goodchild, M., and Gopal, S., 1989. The Accuracy of Spatial Databases. Taylor & Francis.
- Kainz, W., and et. al., 1993. Modeling Spatial Relations and Operations with Partially Ordered Sets, IJGIS. 7(3), 215-229 p.
- Mark, D. J., et. al., 1995. Toward a Standard for Spatial Relations in SDTS and Geographic Information Systems. GIS/LIS'95, Nashville. Tennessee. 686-695 p.
- Peuquet, D. J., and Zhan, C., 1987. An Algorithm to Determine the Directional Relationship between Arbitrarily-Shaped Polygons in the Plane. Pattern Recognition. 20(1), 65-74 p.
- Serra, J., 1982. Image Analysis and Mathematical Morphology. Academic Press.
- Vergin, H., 1994. Integration of Simulation Modeling and Error Propagation for the Buffer Operation in GIS. PE & RS. Vol.60, No.4, 427-435 p.
- Zadeh, L. A., 1965. Fuzzy Sets. Information and Control. vol. 8, 338-352 p.

# Plot Collocation Error: Impacts on Area Estimation

Willem W. S. van Hees<sup>1</sup>

---

**Abstract**--Results of a study conducted to examine area estimation error caused by improper collocation of aerial photo plots and associated ground plots are presented. Analyses considered the complexity of the land cover pattern on the plot, cardinal direction of collocation error, and percentage of area by land cover class on control plots.

During the 1980's, the Forest Inventory and Analysis (FIA)-Anchorage unit of the Pacific Northwest Research Station conducted renewable resource inventories employing a three-phase-with-subsampling inventory design. There were three remotely sensed samples and one ground sample. Regression estimators were developed for several resource quantities, particularly area by land cover class. Results from two inventories showed low correlations between most covariates on the different sampling layers. Statistical consultation indicated inaccurate collocation of sample plots could contribute to poor correlations.

This study focused on the effects of improper collocation between low altitude aerial photo plots and ground plots on estimation of productive forestland area. Results indicate there may be a north-south directional bias and that land cover complexity as estimated by fractal dimension is an important interaction component. Area estimation errors ranged from zero for homogeneous plots to 10.9 percent for plots with more complex land cover patterns.

---

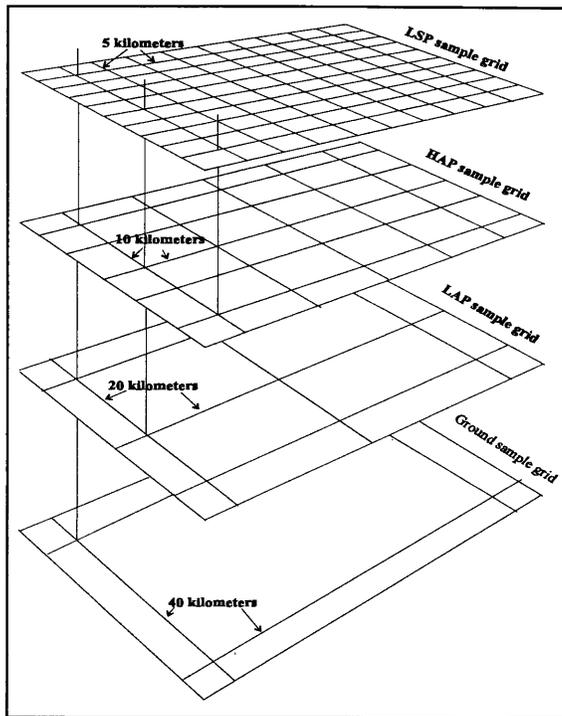
## INTRODUCTION

During the 1980's the Anchorage Forest Inventory and Analysis (FIA) unit of the Pacific Northwest Research Station, conducted experimental multiresource inventories employing a three-phase sampling with subsampling design (Schreuder, Gregoire, Wood 1993). Landsat multispectral imagery (LS), high-altitude color infrared photography (HAP), low-altitude color infrared

---

<sup>1</sup> Research forester, Pacific Northwest Research Station, Anchorage, AK.

photography (LAP), and ground (G) plots were the sampling layers. A grid of 8 ha plots on the LS layer was sampled with successively more extensive subgrids on HAP, LAP, and G. The subgrid of LAP plots was subsampled on the ground. At the LS layer the grid spacing was 5 km, at the HAP layer 10 km, 20 km at the LAP layer, and 40 km on the ground, Figure 1.



**Figure 1--Three-phase sampling with subsampling grid design used for multi-resource inventories in Alaska during the 1980's.**

At each sample location land cover class was evaluated. Land cover polygons were mapped for area estimation at remotely sensed locations and were point sampled at ground locations. Population totals were estimated using several estimators, including regression estimation. Regression results were disappointing in that correlations between covariates and independent variables were not satisfactory. A number of factors, including unrecognized sources of variation and changing objectives were deemed responsible. A possible source of uncontrolled variation was plot collocation. For regression estimators in particular, plots at various sample stages must be accurately collocated. This

study was designed to investigate the hypothesis that apparently minor plot collocation errors can significantly contribute to area estimation error.

Two factors that could contribute to relative error magnitude were considered; the amount of a given land cover class on the plot and the complexity of the land cover pattern on the plot. If a plot is homogenous with respect to land cover class, then small errors in collocation would, at best, result in small relative errors in area estimation. On the other hand, if the plot has small amounts of a given land cover class, then small collocation errors could result in larger relative area estimation errors. To represent land cover class complexity, the fractal dimension of the mapped land cover pattern within the plot was estimated. The hypothesis then, was that area estimation error for a given land cover class (expressed as the difference in percent of plot area in that land cover class on a control plot versus a

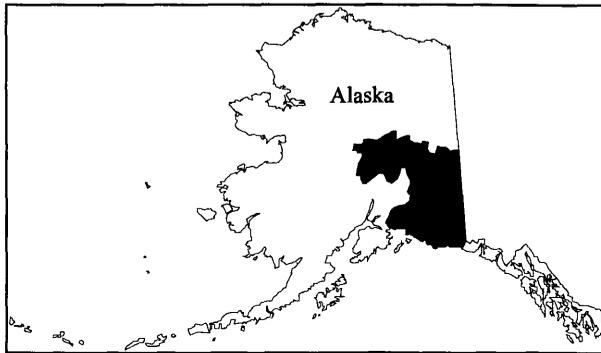
mislocated plot) would be a function of the total plot area in that land cover class (expressed as a percent of total plot area) and of the overall fractal dimension of the land cover pattern on the plot.

## METHODS

### Study Area

Field data for this study did not exist; this study is a simulation in so far as production of collocation errors is concerned. Aerial photos from inventories of the central interior, east-central interior, and south central coastal regions of Alaska (Figure 2.) were used as the data source.

### Data Collection



**Figure 2--Study plots were selected from within the shaded area above.**

Color infra-red aerial photos of 31 different locations were subjectively selected to represent a variety of topographies, latitudes, longitudes, and land cover classes. Aerial photo scales ranged from 1:3,000 to 1:6,000.

An aerial photo interpreter classified and delineated, on transparent overlays, up to 4 land cover classes on each photo. The entire photo was classified. Classes were chosen to simplify the classification process and thereby reduce interpretation error. The classes were productive forestland (forestland capable of producing  $1.4 \text{ m}^3 \cdot \text{ha}^{-1} \cdot \text{yr}^{-1}$ ), other forestland, nonforest, and water.

Each of the resulting 31 land cover class map overlays was digitized. On the digitized images 8 ha control plots were drawn. These plots served as the source of the true, or ground, area. For each control plot, area by land cover class was measured and converted to a relative value ( $P_c$ ).

Improper collocation between these control (or ground) plots and their associated aerial photo plots was modeled by overlaying new plots that were off-center. For each control plot, improperly collocated simulated aerial photo plots were created by overlaying, off-center, new plots moved in one of eight cardinal directions. Directions of movement were randomly chosen without replacement.

The amount of collocation error was approximately 0.65 cm on 1:6,000 scale photos (approximately 40 m on the ground). The magnitude of this error derives from field experience gained during the course of the inventories. Five such mislocated plots were established for each control plot. For each mislocated plot, area by land cover class was also measured and converted to a relative value ( $P_m$ ).

Sets of control and moved plots were created until there were at least 5 observations in each cell of frequency tables for each land cover class. The frequency tables listed number of plots by complexity class (simple vs complex) and percentage of the control plot in the land cover class. The percentages of the control plot in a given land cover class were grouped into quarters. In order to place each plot in a complexity class, estimates of the fractal dimension of the land cover pattern on control plots were made using the dividers method (Sugihara and May 1990). Complexity class was established according to a vegetation complexity model for Alaska (van Hees 1994). Ultimately, 195 sets of control plots and associated moved plots were used for this study.

### Data Analysis

For each set of control and moved plots, the differences in percent of area (or area estimation error), by land cover class, between the moved and control plots were found. The mean of the absolute values of the five differences in percent of area ( $P_e$ ) was then calculated as:

$$P_e = (\sum_i |P_c - P_m|) / 5 \quad i = 1 \text{ to } 5. \quad (1)$$

where:  $P_c$  and  $P_m$  as defined above.

For preliminary analysis, the GLM (General Linear Models) procedure (SAS 1990a) was used to fit a linear regression model, by land cover class, to the data. The model used was:

$$P_e = \beta_0 + \beta_1 (P_c) + \beta_2 (F_d) + \beta_3 (P_c * F_d) \quad (2)$$

where:  $P_e$  = mean absolute area difference,  
 $\beta_0$  = intercept parameter,  
 $\beta_{1,2,3}$  = regression coefficients,  
 $F_d$  = fractal dimension.  $1 \leq F_d < 2$

To investigate the possibility that directional bias existed, the same general linear model as in (2) was used except that individual absolute difference for the

particular direction of mislocation was the dependent variable rather than mean absolute difference over the five mislocated plots.

Subsequent to linear analyses, cluster analyses were conducted using the FASTCLUS procedure (SAS 1990b). Using FASTCLUS, observations are divided into clusters on the basis of Euclidean distances computed from one or more quantitative variables. For this study the variables used for clustering were  $P_e$ ,  $P_c$ ,  $F_d$ , and  $(P_c * F_d)$ .

## RESULTS

Initial screening by linear regression showed significant coefficients for the productive forestland class only.  $T$ -values and analysis of variance statistics are presented in Tables 1 and 2. The predictive value of the regression relation however, as measured by  $r^2$ , was negligible. The  $r^2$  value for this relationship was 0.129.

**Table 1-- $T$ -tests of significance of individual regression coefficients for multiple linear regression analysis of  $P_e$  against  $P_c$ , fractal dimension ( $F_d$ ), and  $P_c * F_d$ .**

Variable	Coefficient	Standard error	$T$	$P$ (two-tail)
Intercept	0.2244931	0.0579952	3.87	0.0002
$P_c$	-0.3542443	0.1019977	-3.47	0.0008
$F_d$	-0.1525531	0.0481601	-3.17	0.0021
$P_c * F_d$	0.3058221	0.0886174	3.45	0.0009

**Table 2--Analysis of variance for multiple linear regression analysis of  $P_e$  against  $P_c$ , fractal dimension ( $F_d$ ), and  $P_c * F_d$ .**

Source	DF	Sum of squares	Mean square	F value	Pr > F
Model	3	0.00942545	0.00314182	4.15	0.0085
Error	84	0.06352763	0.00075628		
Total	87	0.07295308			

Adding a squared interaction term so that the model became

$$P_e = \beta_0 + \beta_1 (P_c) + \beta_2 (F_d) + \beta_3 (P_c * F_d) + \beta_4 (P_c * F_d)^2, \quad (3)$$

improved predictive power. The  $r^2$  value for the expanded model was 0.375. Again, the productive forestland class was the only class for which all regression coefficients were significant.

The mean difference ( $P_e$ ) between the percentages on the control plots versus the mislocated plots ranged from zero to 10.9 percent with a mean of 4.3 percent and

standard deviation of 2.9 percent. In the case where there was zero difference between the control and the mislocated plots all plots were homogenous. Land class pattern fractal dimension ( $F_d$ ) ranged from 1.000 for homogenous plots to 1.530 for the most complex patterns. Mean  $F_d$  was 1.165 with a standard deviation of 0.109.

Model (3) was used to examine the data for possible directional bias. That is, could the direction in which mislocation of the aerial photo plot occurred affect the magnitude of relative area estimation error. Table 3 presents the results of this analysis. Noticeable is the strength of the estimated regressions for those observations where the direction of mislocation was along the north/northeast by south/southwest axis (highlighted figures in Table 3) when compared with the estimates for other directions.

**Table 3--F-values and T -scores for estimated coefficients of model (3) by direction of movement of improperly collocated plots for productive forestland.**

Direction	df	F	T for $H_0 = 0$				$r^2$
			$P_c$	$F_d$	$P_c * F_d$	$(P_c * F_d)^2$	
North	4,54	<b>5.82</b>	<b>-2.99</b>	-2.70	<b>3.70</b>	<b>-4.37</b>	<b>.3011</b>
Northeast	4,42	<b>4.02</b>	-2.59	-2.61	<b>3.30</b>	-1.83	<b>.2767</b>
East	4,57	2.93	-1.92	-1.32	2.53	<b>-3.29</b>	.1706
Southeast	4,38	1.53	-1.81	-1.21	2.28	-1.68	.1388
South	4,60	<b>6.12</b>	<b>-3.31</b>	-2.53	<b>4.04</b>	<b>-3.96</b>	<b>.2898</b>
Southwest	4,30	<b>4.09</b>	-1.74	-1.80	2.83	-2.87	<b>.3526</b>
West	4,64	1.44	-1.77	-1.50	2.01	-2.08	.0824
Northwest	4,35	1.38	-1.12	-0.77	1.08	-1.16	.1359

Clustering of the data was undertaken to examine the possibility that the predictive power of the model (using equation (1)) could be improved. FASTCLUS was run setting the number of clusters at 3 and 4. Overall  $r^2$  for the 4-cluster grouping was higher than for the 3-cluster grouping (0.886 vs 0.829). Clusters were separated, by FASTCLUS, into the following ranges of  $P_c$  with almost no overlap. The groups of  $P_c$  were 0 - 15% (n=18), 15% - 45% (n=20), 45% - 75% (n=31), and 75% - 100% (n=19). Model (2) was then used to examine the within cluster regression relationships. In the first and last clusters the regression relationships were quite strong ( $r^2 = 0.729$  and  $0.841$  respectively). The relationship for the second cluster was much weaker ( $r^2 = 0.385$ ) and for the third cluster the regression relationship was essentially nonexistent ( $r^2 = 0.026$ ).

## DISCUSSION

With regards to the productive forestland component of the study area, the hypothesis that area estimation error would be a function of the percentage of the

control plot ( $P_c$ ) in a land cover class and the fractal dimension of the land cover pattern is supported by the results of this study.

Although this study does not provide substantive clues as to why the model was more informative for productive forestland than for other land cover classes a possible explanation lies in the physiography of the study area and the discovery of possible directional bias. Patterns of productive forestland tend to be linear in the study area. In interior Alaska productive forestland is found along rivers on south facing slopes whereas in coastal regions it is “marginal” - that is along the margins next to water and up to about 600 m above sea level. Rivers and water bodies in the study area have a predominantly east/west orientation. Thus, plot mislocations to the north or south would likely have more impact on error magnitude than mislocations to the east or west.

Clustering of the observations did improve predictive ability for certain ranges of the data; particularly the upper and lower ends of the control plot percentage scale. It would be possible then, to more accurately assess impacts of mislocation knowing the percentage of the plot in a land cover class along with an estimate of fractal dimension of the mapped land cover classification.

Although the results of this study do not provide strong modeling capabilities they do provide useable guidelines. It is apparent that small mislocations between aerial photo and ground plots can produce significant area estimation errors (up to 10 percent) and that inventory designers and data analysts must be aware of physiography in order to consider possible introduction of directional bias. Also, there may be some predictive capability for extremes of land cover percentages.

#### LITERATURE CITED

- SAS. 1990a. SAS/STAT User's guide, Version 6, Fourth ed., vol. 2. SAS Institute, Inc., Cary, NC. USA.
- SAS. 1990b. SAS/STAT User's guide, Version 6, Fourth ed., vol. 1. SAS Institute, Inc., Cary, NC. USA.
- Schreuder, Hans T., T. G. Gregoire, and G. B. Wood. 1992. Sampling methods for multi-resource forest inventory. John Wiley and Sons, Inc., 605 Third Ave., New York, New York. 10158-0012. 446 p.
- Sugihara, George & May, Robert M. 1990. Applications of fractals in ecology. Trends in Ecology and Evolution, Elsevier Science Publishers Ltd., (UK), 5(3): 79-86.
- van Hees, Willem W. S. 1994. A fractal model of vegetation complexity in Alaska. Landscape Ecology. 9(4):271-278.

# Sensitivity of Landscape Pattern Metrics to Misclassification and Differences in Land Cover Composition<sup>1</sup>

James D. Wickham<sup>2</sup>, Robert V. O'Neill<sup>3</sup>, Kurt H. Riitters<sup>2</sup>,  
Timothy G. Wade<sup>4</sup>, and K. Bruce Jones<sup>5</sup>

---

**Abstract.** Calculation of landscape metrics from land cover data is becoming increasingly common. Some studies have shown that these measurements are sensitive to differences in land cover composition, but none are known to have tested also their sensitivity to land cover misclassification.. An error simulation model was written to test the sensitivity of selected landscape pattern metrics to misclassification, and regression analysis was used to determine if these metrics were significantly related to differences in land cover composition. Comparison of sensitivity and regression results suggests that differences in land cover composition need to be about 5 percent greater than the misclassification rate to be confident that differences in landscape metrics are not due to misclassification.

---

---

1 Forthcoming in *Photogrammetric Engineering & Remote Sensing*, 1996, Volume 62.

2 Tennessee Valley Authority, Norris, TN 37828.

3 Oak Ridge National Laboratory, Oak Ridge, TN 37831.

4 Desert Research Institute, Reno, NV 89512.

5 U.S. Environmental Protection Agency, Las Vegas, NV 89119.

# Spatial (In)consistency of Watershed Delineations Among Agencies and Scales in Pennsylvania

Wayne L. Myers<sup>1</sup>, Barry M. Evans<sup>2</sup> & Michael C. Anderson<sup>3</sup>

---

**Abstract.**--The Office for Remote Sensing of Earth Resources at Penn State University serves as a GIS resource center and repository for the Pennsylvania Department of Environmental Protection. One of the tasks in this context involves harmonizing different map perspectives on watersheds between agencies and scales, such as USGS and the State Water Plan. This provides an interesting opportunity to study the nature, degree, and distribution of spatial disparity in mappings of supposedly well-defined and naturally hierarchical environmental elements. Spatial and statistical patterns of discrepancy in divides and pour points are considered relative to scale and environmental resource management focus.

---

## INTRODUCTION

The Office for Remote Sensing of Earth Resources (ORSER) in the Environmental Resources Research Institute (ERRI) at The Pennsylvania State University has a considerable history of spatial approaches to watershed studies and hydrologic modeling. ORSER also serves as a GIS resource center and spatial data repository for the Pennsylvania State Department of Environmental Protection (DEP). The latter agency is regularly concerned with water resource issues ranging in scope from local up to multi-state concerns regarding Chesapeake Bay. These multifarious water resource purviews involve several mappings of watersheds at different levels of discrimination, originating from different authorities. Although one usually thinks of watersheds as being physiographically determined, some degree of inconsistency among the suite of mappings was evident. In the interest of

---

<sup>1</sup>*Assoc. Prof. of Forest Biometrics, Penn State Univ., Univ. Park, PA.*

<sup>2</sup>*Senior Research Assistant, Environ. Resources Research Inst., Penn State Univ., Univ. Park, PA.*

<sup>3</sup>*Research Technician, Environ. Resources Research Inst., Penn State Univ., Univ. Park, PA.*

clarity and consistency, ORSER was tasked by DEP to harmonize the several mappings in a GIS framework.

Comparative analysis of habitat variables among and within watersheds is one of the undertakings for Pennsylvania's ongoing Gap Analysis of conservation status for vertebrates. A Gap successor project will use digital elevation models and stream reach files to partition prior watershed mappings into "reachsheds" which will then be reaggregated according to similarity criteria as a basis for critical area assessment of landscapes. Because of these research dependencies on accuracy and consistency of watershed files, it was decided to conduct some accuracy investigations in conjunction with the harmonization effort.

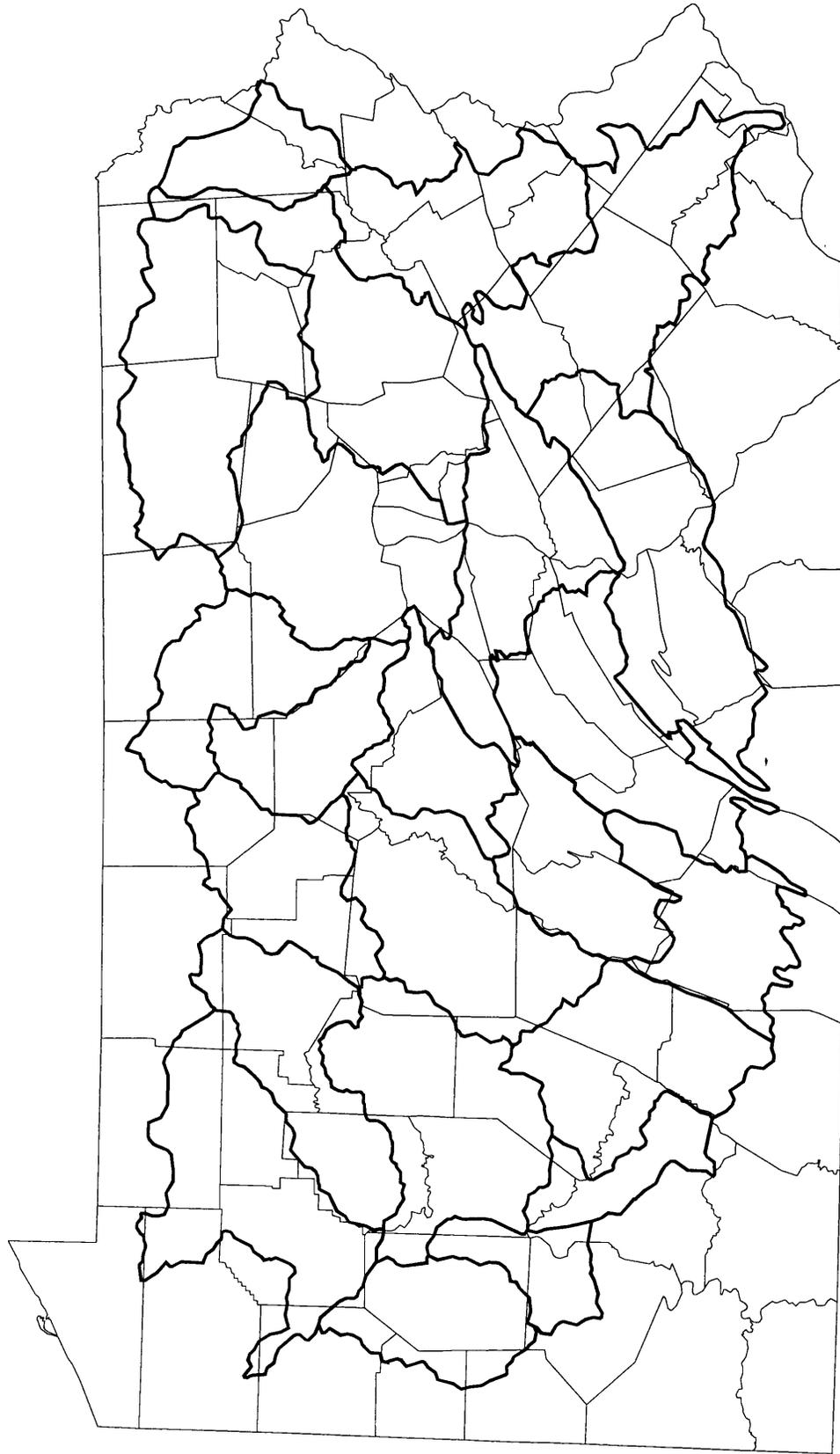
### **PROCEDURE**

Three different watershed data series were selected for the present comparative analysis. USGS digital drainage basin boundaries of named streams in Pennsylvania (Hoffman & Bubb, 1995) comprised the lower tier of (smallest) watersheds. Watersheds recognized by the State Water Plan of the Pennsylvania Bureau of Water Resources Management (1:250,000 scale) constituted the middle tier. There are a total of 104 such watersheds delineated for the state. USGS 8-digit hydrologic unit boundaries (1:250,000 scale) represented the top tier of (largest) watersheds. Analysis was restricted to top-tier watersheds lying entirely within the boundaries of the Commonwealth. Outlines of these 23 watersheds are shown in Figure 1.

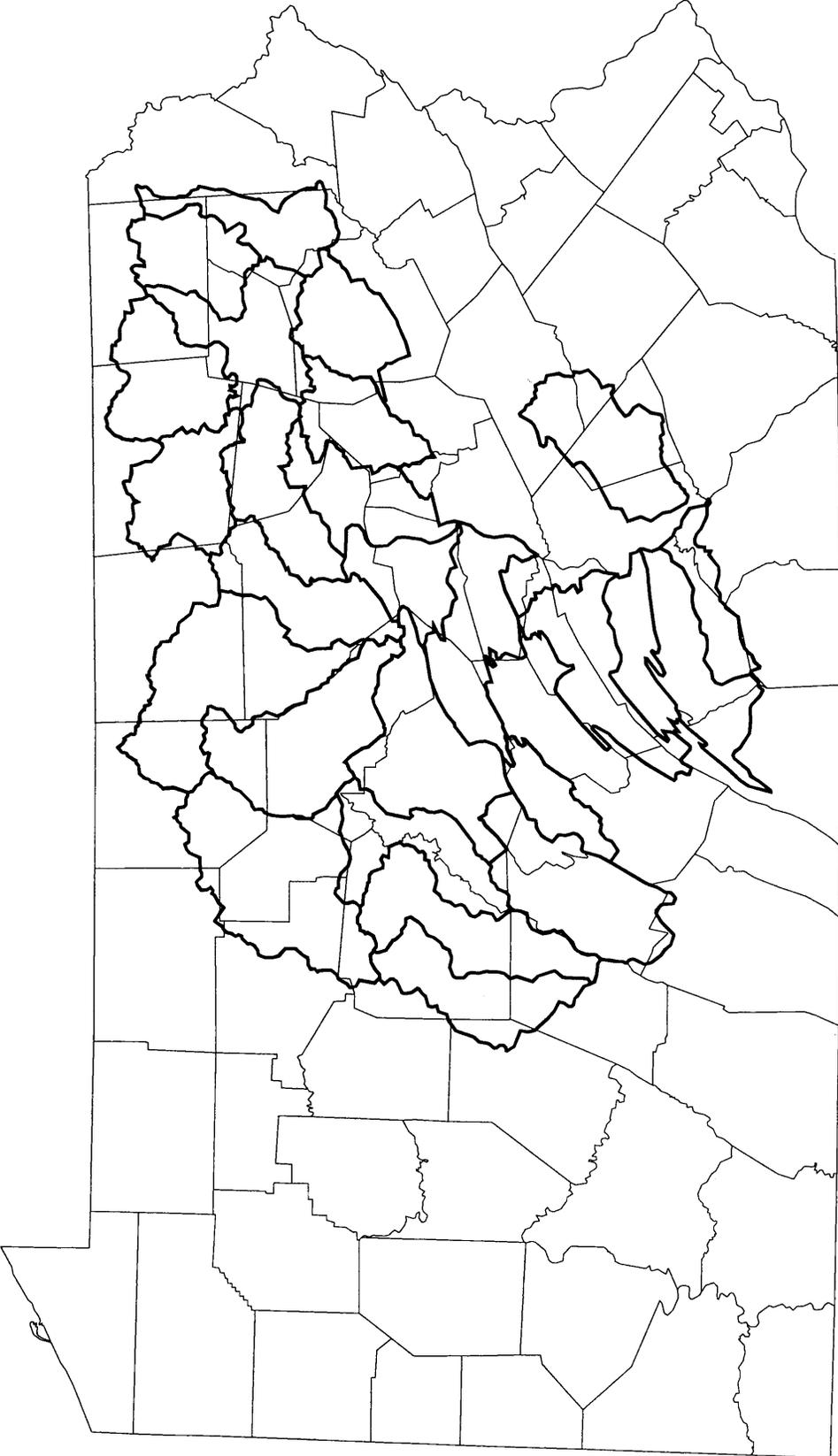
The 26 middle-tier watersheds comprising the Susquehanna River Basin were used in the first stage of analysis. Outlines of these 26 watersheds are shown in Figure 2. The 3,585 lower-tier watersheds for the Susquehanna Basin (not shown) were aggregated upward to match the corresponding 26 middle-tier watersheds. Matched watersheds were then compared for discrepancies in boundaries and pour points using Arc/Info software.

All of the whole watersheds for both coverages (smaller scale and aggregated) were extracted and placed in their own coverages. They were both numbered the same way, i.e., watershed-id #1 represents the same watershed in both coverages. The mean area and mean perimeter of each pair matched by id# were then computed. A union operation was then performed on the two coverages. The number of slivers and sliver area were determined for each pair. An index of sliver width was developed as the sliver area divided by the average perimeter.

**Figure 1. USGS Watersheds**



**Figure 2. State Water Plan Watersheds**



For automated determination of pour points, nodes were placed at 100 meter intervals along the boundaries, and elevations transferred from a corresponding 100-meter DEM. The DEM was a 3 arc-second USGS DEM resampled on an Albers equal-area projection used for the watersheds. Pour points were determined as the minimum elevation on the boundary.

In the second stage of analysis, middle-tier watersheds were similarly aggregated to match the 23 top-tier watersheds. Metrics of discrepancy were determined in the same manner as for the first stage of analysis.

## RESULTS

Table 1 contains detailed results for aggregating from smallest watersheds to mid-level watersheds.

**Table 1. Results of aggregating to mid-level (from smallest) watersheds.**

<b>ID</b>	<b>Pour pt. dist(m)</b>	<b>Avr. perim. (km)</b>	<b>Avr. area (ha)</b>	<b>No. of slivers</b>	<b>Sliver area(ha)</b>	<b>Sliver width(m)</b>
1	36,441	211.5	142,920	95	7,198	340
2	2,815	344.7	253,958	195	10,419	302
3	3,653	214.0	122,677	103	8,002	374
4	165	186.4	106,361	84	7,928	425
5	431	231.9	93,026	89	11,040	476
6	558	285.4	133,025	138	10,362	363
7	344	317.7	269,886	170	10,901	343
8	561	279.5	202,310	146	8,013	287
9	3,033	240.6	127,934	134	8,658	360
10	4,828	242.7	134,558	121	10,324	425
11	14,329	255.6	116,963	112	10,624	416
12	887	182.3	105,931	58	7,752	425
13	604	187.1	99,741	102	5,039	269
14	1,997	212.4	90,732	45	12,731	599
15	330	263.9	124,220	119	9,513	361
16	561	295.4	199,776	103	9,866	334
17	5,295	293.6	161,010	122	14,447	492
18	5,124	243.8	129,307	110	12,562	515
19	1,136	294.9	151,060	96	10,608	360
20	453	259.9	192,572	108	9,602	369
21	861	170.2	63,936	62	5,434	319
22	302	291.6	145,584	134	6,962	239
23	683	215.7	148,275	99	6,012	279
24	2,162	198.3	77,829	68	5,461	275
25	3,613	265.3	131,006	97	9,079	342
26	11,877	184.2	64,804	67	9,577	520

Table 2 contains results of aggregating to highest level (largest) watersheds from mid-level watersheds.

**Table 2. Results of aggregating to highest level (largest) watersheds.**

<b>ID</b>	<b>Pour pt. dist(m)</b>	<b>Avr. perim. (km)</b>	<b>Avr. area (ha)</b>	<b>No. of slivers</b>	<b>Sliver area(ha)</b>	<b>Sliver width(m)</b>
1	191	423.2	512,086	107	33,934	802
2	2,336	294.2	253,699	83	11,990	408
3	4,868	442.9	436,231	79	29,153	658
4	140	500.2	463,004	116	26,806	536
5	719	198.6	151,596	38	11,749	592
6	926	280.7	271,222	66	15,189	541
7	1,942	344.1	322,189	52	22,876	665
8	1,723	246.5	202,984	59	11,144	452
9	140	379.3	471,123	101	15,941	420
10	1,066	217.4	218,444	63	10,243	471
11	5,537	366.9	437,091	68	28,285	771
12	2,939	389.2	353,314	92	19,726	507
13	926	327.5	416,670	68	15,792	482
14	1,723	265.1	199,032	51	11,132	420
15	920	429.3	375,489	99	20,543	479
16	677	198.3	127,887	36	14,501	731
17	2,079	530.2	496,632	123	32,370	611
18	628	271.4	256,782	54	14,137	521
19	920	400.6	377,071	82	23,534	587
20	7,532	297.2	351,141	43	24,995	841
21	5,145	440.5	478,482	73	39,309	892
22	403	228.5	133,976	42	20,869	913
23	455	340.0	246,702	53	25,012	736

The area in slivers after aggregation reflects the uncertainty of boundary delineation between scale levels for the watersheds. Pooling the sliver areas and mean areas over all watersheds, the sliver area represents 6.6% of the mean area. The corresponding percentage for aggregating from mid-level to large watershed level is 6.3%. These percentages are surprisingly consistent, and thus indicative that the 6% figure holds across scales for Pennsylvania watershed demarcation.

The observed distance between aggregated and original pour points is quite another matter. This distance differential is highly variable, ranging from less than 200 meters to more than 7 km for scaling up from mid-level to large watersheds. The variability is even greater for scaling up small watersheds to mid-level watersheds.

Upon investigation, the 36+ km figure for watershed id#1 in the latter set constitutes an anomaly. In this particular case, the pour points before and after aggregation were placed at opposite ends of the watershed. This is presumably attributable to inconsistency of the DEM used for pour point location.

Irregularities in the DEM undoubtedly also contribute to large pour point discrepancies in the remaining cases, but other factors are also involved. The sliver width index is one measure of mean sliver width. This is on the order of one-half kilometer, being somewhat less for the lower aggregation and somewhat more for the upper aggregation. This would account for a corresponding order of difference in pour points.

Constrained reaches are particularly problematic relative to location of pour points. Such reaches give rise to "corridor" watersheds having very limited area, but causing substantial displacements of pour points depending on whether or not they are treated separately. There are also more arbitrary instances such as cases where delineations track parts of boundaries for civil divisions.

## **CONCLUSIONS**

Multi-scale analysis involving spatially explicit datasets from different sources at different levels of detail will require that inconsistencies be harmonized across scales. The process of harmonization affords opportunity for comparative analysis of inconsistencies to determine relative orders of spatial accuracy.

Comparative analysis of watershed delineations at three scales in Pennsylvania reveals that differing positions for drainage divides induce errors corresponding to approximately 6% of watershed area. In contrast, the range of difference in pour point locations spans more than an order of magnitude.

## **REFERENCES**

Hoffman, S. A. and J. E. Bubb. 1995. Digital drainage basin boundaries of named streams in Pennsylvania. U.S. Geological Survey, Open-File Report 95-xxx, Lemoyne, PA.

# COMMUNICATING THE RESULTS OF ACCURACY ASSESSMENT: METADATA, DIGITAL LIBRARIES, AND ASSESSING FITNESS FOR USE

Michael F. Goodchild<sup>1</sup>

---

The concept is proposed of a data life cycle extending from the initial collection of data in the field or by remote sensing to eventual archiving. Although GIS has been perceived to date largely as an analytic technology confined to a small part of the data life cycle, its influence is increasingly seen as extending throughout the cycle, in the form of newer technologies such as field GIS and digital spatial data libraries. The concept of a data model is introduced and examples are given of the role of data modeling along the life cycle. Accuracy assessment and concern for broader issues of data quality are critical if data are to be assessed as to fitness for use by the various actors in the life cycle. GIS data models must be extended to accommodate essential information on accuracy. The role of accuracy assessment in metadata is discussed, and a new view of metadata is proposed with extended functions and a hierarchical structure.

---

---

<sup>1</sup> *Professor of Geography and Director, National Center for Geographic Information and Analysis, University of California, Santa Barbara, CA 93106-4060. Email: good@ncgia.ucsb.edu*

# Covariate-Directed Sampling for Assessing Species Richness

G.P. Patil, Glen Johnson  
and Matteo Grigoletto<sup>1</sup>

---

**Abstract.**—Since species richness is a spatially non-additive variable, it can not be estimated with conventional moment estimators. We may, however, exploit the species-area relationship which implies that the number of species increases according to a law of diminishing returns as the sampled area increases. An efficient sampling method would then maximize the number of species encountered within a fixed amount of area that can be affordably sampled. We suggest that when spatial covariate information is available, it should be exploited for directing the location of sample units in such a manner that increases the habitat diversity observed within a minimum number of sample units. This approach was evaluated through a retrospective assessment of breeding bird species richness in Pennsylvania using cumulative tree richness as the covariate. Working with EMAP hexagons (635 square kilometers) as the primary sampling unit, we found that although tree richness was a fairly weak covariate, it still outperformed random sampling.

---

## INTRODUCTION

With such a strong and legitimate concern for the alarming loss of biodiversity on our planet (Wilson, 1988; Stevens, 1995), monitoring methods are essential for quantifying the biodiversity of large geographic regions. With all the shortcomings of “indices”, actual species richness (the number of different species) appears to be the least controversial measure of biodiversity. Meanwhile, biodiversity researchers recognize that reliable methods of estimation still require development (Yoon, 1995).

Since species richness is a spatially non-additive variable, we can not estimate its total from conventional moment estimators, such as by multiplying a sample mean by the number of population units in the sampling frame. For this reason, we turn to species-area curves, which are used by ecologists for several reasons (Kilburn, 1966), including the prediction of species richness in larger areas than those sampled (Evans, Clark and Brand, 1955).

---

<sup>1</sup>Center for Statistical Ecology and Environmental Statistics, Department of Statistics, Penn State University, University Park, PA 16801

The species-area relationship basically states that as the area within a homogeneous habitat increases, the number of different species encountered will also increase until “a point of no return”, after which increasing the area does not further increase the number of different species encountered. The common model for this process, as originally proposed by Arrhenius (1921), is a power function, presented as  $S = kA^z$ , where  $S$  is the species richness and  $A$  is the area, while  $z$  and  $k$  are population specific parameters. For applications to wildlife conservation issues, see Usher (1985). Although the power function has been traditionally used to model the species-area relationship within homogeneous habitat, Johnson and Patil (1995) observed a classical power function response for breeding bird richness across the whole state of Pennsylvania which encompasses very heterogeneous habitat. If the power function is fit from sample data, its ability to extrapolate to a larger area is limited by upward bias since this model is unbounded (Williams, 1995).

Our objective is to develop a sampling plan that maximizes the acceleration of a species-area curve towards its plateau in order to encounter the most species within a sampled area. This is especially critical when sampling from a very large geographic area like the state of Pennsylvania since this can rapidly become a very expensive exercise.

## SAMPLING STRATEGIES

When constructing a species-area curve from successive aggregation of discrete sample units, the usual approaches are to either combine sample units in a continuous fashion or to combine units that are obtained at random from throughout the region of interest.

When habitat is diverse across the region, random aggregation may result in a steeper curve than is obtained from continuous aggregation because spatially discontinuous sample units may encounter more diverse habitats, therefore increasing the chance of encountering different species. The expected number of species encountered in  $n$  sample units of equal size obtained at random,  $E[S_n]$ , can be readily computed (Kobayashi, 1979) for providing a benchmark to compare other sampling protocols.

An alternative approach to continuous aggregation or random sampling is to perform directed sampling based on values of some covariate that are readily available for the sample units. With the advent of geographic information systems, we feel that such covariate information is becoming more readily available for geographic areas at the landscape scale and above. The desired property of a covariate would be to direct sampling in a manner that accelerates the species-area curve faster than would be observed with spatially continuous or random sampling.

## BREEDING BIRDS IN PENNSYLVANIA

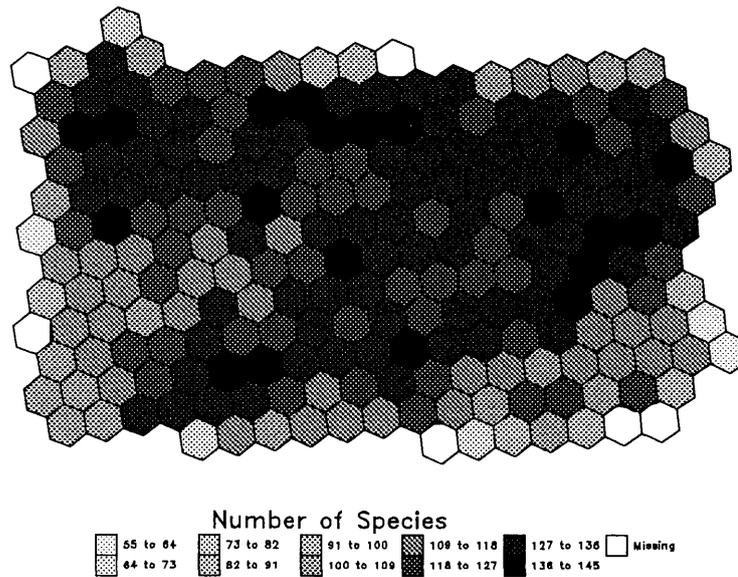


Figure 1: **Bird richness in the hexagons.**

We evaluated the proposed approach of covariate-directed sampling through a retrospective study of a known community of Breeding birds in Pennsylvania. Our database is described in Johnson and Patil (1995). The sampling frame consists of a tessellation of Pennsylvania by hexagons, each 635 km<sup>2</sup>, of the Environmental Protection Agency’s Environmental Monitoring and Assessment Program (EMAP). Associated with each hexagon are species lists for breeding birds, other vertebrate groups and trees. While other groups are based on records of occurrence, the breeding birds are based on the much more thorough Pennsylvania Breeding Bird Survey (Brauning, 1992). The distribution of species richness for breeding birds with respect to EMAP hexagons is displayed in Figure 1 in the form of a greyscale thematic map.

Of the information available in our database, tree species presented the most promising covariate for choosing an optimal hexagon ordering for ultimately measuring bird species richness. We basically hypothesized that differences in bird species are likely to be associated with differences in tree species; therefore, if hexagons are chosen in an order that corresponds to maximum acceleration of the tree species richness curve, using this same ordering will accelerate the bird richness curve. Constructing the optimal tree species richness curve was performed by choosing the first hexagon as the one containing the highest tree species richness. After noting which species were in the first hexagon, all members of these species were deleted from the remaining hexagons. The second hexagon was then chosen as the one containing the highest tree richness. Steps 1 and 2 were then repeated until all the tree

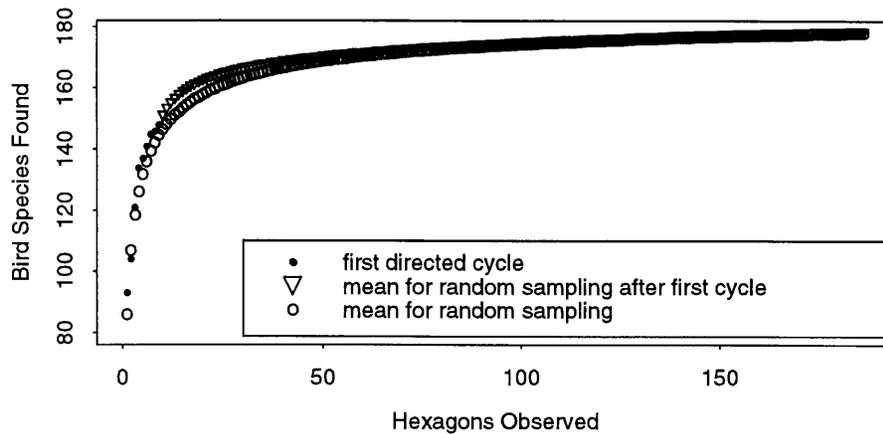


Figure 2: Species-area curve from tree-directed sampling for one cycle, followed by the expected value from subsequent random sampling. The expected curve from completely random sampling is also shown.

species had been accounted for.

We discovered that all tree species were accounted for within the first nine hexagons sampled. At this point we experimented with two techniques. The first one is to randomly sample part of the remaining hexagons. The second technique is to reintroduce all of the tree species that are within the unsampled hexagons, repeating the procedure used for the first cycle, and so on for a certain number of cycles (this is called a completely directed procedure). At each step the new bird species found in the hexagon were recorded. The completely directed procedure proved to perform somewhat better, as seen in Figures 2 and 3 which provide results over the whole state.

For the protocol which sampled the first nine hexagons by tree-directed sampling, we fit the power function model, via linear regression, for additional random samples of 20, 30 and 50. The results are presented in Table 1, where extrapolation to the total area (statewide) appears to yield unacceptably high bias. Since we are using a covariate (number of tree species) for the first cycle followed by random sampling, then the number of new tree species considered increases just for the first nine hexagons, and it becomes constant thereafter (just the area increases). To take this into account, we might use a model of the form  $S = kA^zT^p$ , where  $T$  is the number of tree species and  $p$  is a new parameter. Such a model has been used by Rafe *et al.* (1985); what they call *habitat heterogeneity* in our case is represented by number of tree species. The extrapolated curves for both models are compared in Figure 4 for the sampling protocol of one tree-directed cycle, followed by 20 randomly chosen hexagons.

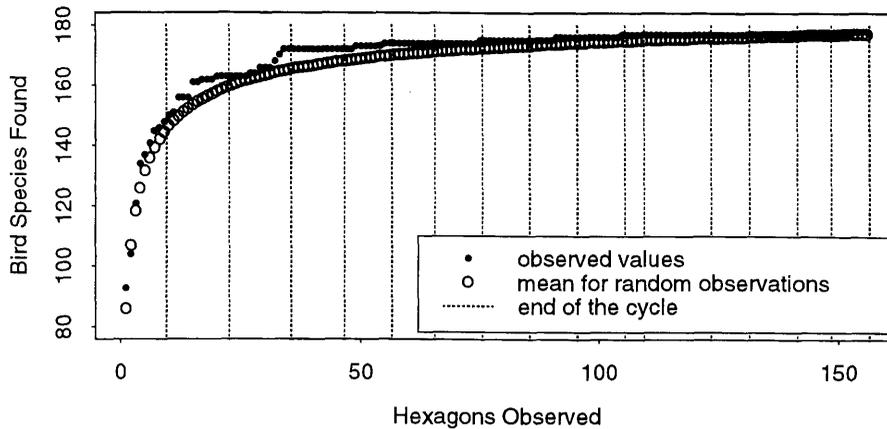


Figure 3: Species-area curve from completely tree-directed sampling, with demarcation of each sampling cycle. The expected curve from completely random sampling is also shown for comparison.

Table 1: Extrapolated species richness (EX) for one tree-directed cycle (9 hexagons) plus subsequent random samples of size 20, 30 and 50. Bias equals the extrapolated minus the total known species.

Sample size	$\hat{k}$	$\hat{z}$	EX	bias
29	36.87	0.157	230	+52
39	42.76	0.139	216	+38
59	51.42	0.117	202	+24

Here we see a substantial reduction in bias from incorporating the covariate.

The main drawback of a model which incorporates the covariate is that there is no trivial way to use it when covariate-directed selection of *all* sample units comes into play (completely directed procedure). There are no clear values of the variable  $T$  to be used in the second and the following cycles. This problem is important since we obtain the steepest species-area curve with the completely directed procedure.

Besides tree species, we also had available information on (i) fish, (ii) mammals, (iii) reptiles and amphibians and (iv) butterflies and skippers. Using these covariates we reapplied the same analysis as with trees, but did not discover a stronger covariate than tree richness. Results from studying these other covariates can be found in Grigoletto, *et al.* (1995).

## DISCUSSION

Most of the cycle lengths in Figure 5 vary little around 10 hexagons. Why

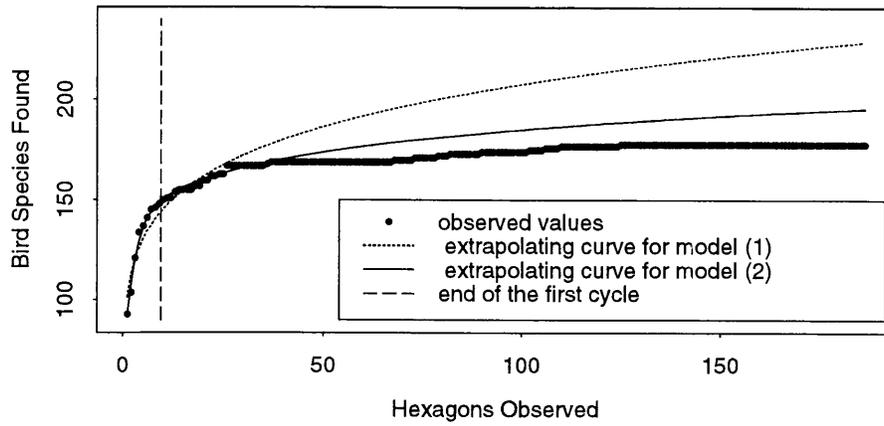


Figure 4: Extrapolating curves for Model 1 ( $S = kA^z$ ) and Model 2 ( $S = kA^zT^p$ ).

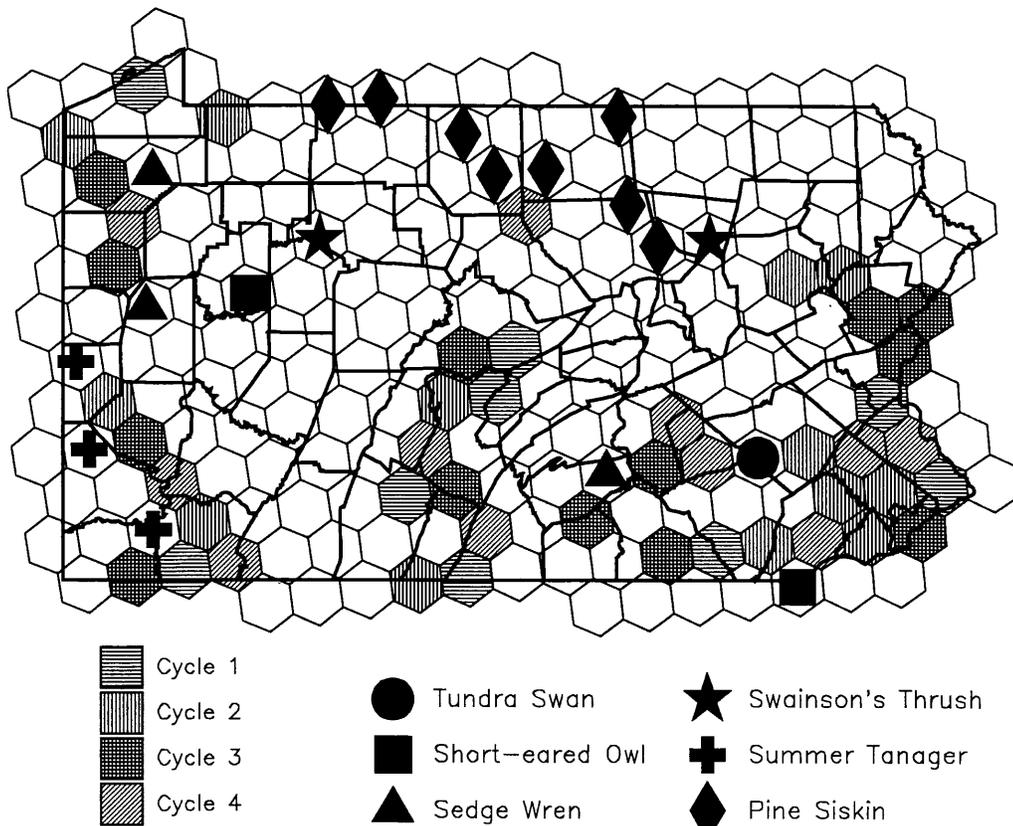


Figure 5: Observed hexagons and unencountered species after 4 cycles.

does that happen? If we look at it is clear that the observed sites tend to cluster. Since we are using tree-directed sampling, this happens because if there are certain areas consisting of multiple hexagons with a high number of tree species, then in different cycles we tend to re-observe these areas. In fact, the larger clusters do contain hexagons from each of the four cycles.

After four cycles, only six bird species remain unencountered. In Figure 5, we see that most of the unencountered species are rare and appear just in one, two or three hexagons. Pine Siskin is an exception since this bird is quite spread over central northern Pennsylvania. Since this region is a mostly forested area, this seems to suggest that if the "area covered by forest" could be used jointly with tree richness, we would have a stronger covariate.

The primary purpose of this paper is to suggest the idea of covariate-directed sampling for achieving observational economy when sampling to estimate species richness, a spatially non-additive variable. We illustrated the approach with a retrospective study of breeding birds in Pennsylvania, where the database at our disposal provided species lists for several other communities. Of these, tree richness appeared to be the best available covariate. Although tree richness was not strongly correlated with bird species richness, tree-directed sampling did accelerate the bird species-area curve faster than was expected with completely random sampling. Therefore, if other covariates can be obtained that are more strongly correlated with bird richness, then such covariates are expected to further improve the sampling efficiency.

An outstanding question that still remains is how to *estimate* species richness statistically, in a manner that allows construction of confidence bounds. Progress has been made (Bunge and Fitzpatrick, 1993) for the case when population densities are also measured within each sampled species; however, we were faced with presence/absence responses for each species, where the data-analytic approach appears to be the only alternative.

### Acknowledgments

Prepared with partial support from the United States Environmental Protection Agency, Environmental Monitoring and Assessment Program, EMAP Design and Statistics Group under a Cooperative Agreement Number CR-821783. The contents have not been subjected to Agency review and therefore do not necessarily reflect the views of the Agency and no official endorsement should be inferred.

### REFERENCES

- Arrhenius, O. 1921. Species and area. *Journal of Ecology*, 9:95–99.
- Brauning, D.W. 1992. *Atlas of Breeding Birds in Pennsylvania*. University of Pittsburgh Press, Pittsburgh. 484 pp.

- Bunge, J. and Fitzpatrick, M. 1993. Estimating the number of species: a review. *Journal of the American Statistical Assoc.*, 88:364-373.
- Evans, F.C., Clark, P.J. and Brand, R.H. 1955. Estimation of the number of species present on a given area. *Ecology*, 36:342-343.
- Grigoletto, M., Johnson, G., Patil, G.P. and Taillie, C. 1995. Using Covariate-Directed Sampling of EMAP Hexagons to Assess the Statewide Species Richness of Breeding Birds in Pennsylvania. Technical Report no. 95-1102. Center for Statistical Ecology and Environmental Statistics, Department of Statistics, Penn State University, University Park, PA.
- Johnson, G.D. and Patil, G.P. 1995. Estimating statewide species richness of breeding birds in Pennsylvania. *Coenosis*, 10(2-3):81-87.
- Kilburn, P.D. 1966. Analysis of the species-area relation. *Ecology*, 47:831-843.
- Kobayashi, S. 1979. Species-area curves. in Ord, J.K., Patil, G.P. and Taillie, C (eds.), *Statistical Distributions in Ecological Work*, pp. 349-368. International Co-operative Publishing House, Fairland, Maryland.
- Rafe, R.W., Usher, M.B and Jefferson, R.G. 1985 Birds on reserves: the influence of area and habitat on species richness. *Journal of Applied Ecology*, 22:327-335.
- Stevens, W.K. 1995. How many species are being lost? Scientists try new yardstick. *The New York Times*, p. C4, July 25.
- Usher, M.B. 1985. Implications for the species-area relationship for wildlife conservation. *Journal of Environmental Mngt.*, 21:181-191.
- White, D., Kimerling, A.J. and Overton, W.S. 1992. Cartographic and geometric components of a global sampling design for environmental monitoring. *Cartographic and Geographic Information Systems*, 19(1):5-22.
- Williams, M.R. 1995. An extreme-value function model of the species incidence and species-area relations. *Ecology*, 76(8):2607-2616.
- Wilson, E.O. 1988. *Biodiversity*. National Academy Press, Washington, D.C., 521 pp.
- Yoon, C.K. 1995. Monumental inventory of Costa Rican forest's insects under way. *The New York Times*, p.C4, July 11.

# Data Accuracy to Data Quality: Using spatial statistics to predict the implications of spatial error in point data

**Adam Lewis and Michael F. Hutchinson**

---

**Abstract.**- Data error has received a good deal of attention, however, data error alone is not very informative. To judge the value of a dataset for a specific application, measures of data *quality* are needed.

A quantitative approach to the problem of estimation of data quality for point data is presented, based on the use of geostatistics to characterise the GIS datasets with which the points are to be overlaid.

Results demonstrate that the intended application of data is critical to the assessment of data quality, and that the suitability of data for a given application can be assessed with only a basic model of the absolute spatial error associated with the point data.

---

## INTRODUCTION

A substantial body of literature exists on the subject of error in spatial databases.

For point data, spatial error can be modelled as a random process. Most simply a bivariate normal distribution can be used, but other symmetric probability density functions have also been used, eg. Bolstad *et al.* (1990). The scaling parameter of the distribution, such as the standard deviation, may be referred to as *epsilon*. An alternative to the probabilistic model is the deterministic interpretation, in which all points are considered to fall within a distance, *epsilon*, of the true location. This extends to the *epsilon-band* model of spatial error in linear features. These interpretations of spatial error are often quoted as a standard in mapping organisations (Goodchild 1988), for instance "90% of points are within 25 metres of the correct location".

Attribute errors arise because thematic maps, DEMs and other spatial datasets are approximations. The polygon representation of spatial variability maps homogeneous units with distinct boundaries, however it is well understood that for natural resource data these polygons are heterogeneous. Where continuous spatial variation is represented on a grid or lattice, as with a DEM, there is a residual error in the model, which may be expressed as the RMS error.

Both spatial and attribute errors are, typically, spatially autocorrelated; the error in one location is dependent on the errors nearby. This reduces the consequences of error in some circumstances - digital contours do not cross randomly although they may be within each others *epsilon-band* - but complicates the modelling of errors. The probabilistic point-*epsilon* model of spatial errors cannot be extended to lines without complication, and models of attribute error require simulation of autocorrelated processes (Goodchild *et al.* 1992).

## Data error versus data quality

Few studies have focussed on the practical implications of errors in spatial data. Ultimately the question is not the absolute error associated with a dataset, but whether the datasets available are of adequate *quality* to support a given *use*.

This paper develops the relationship between *locational accuracy*, and *data quality*, the ultimate aim being to provide a basis from which to make informed statements on accuracy.

## DATA SOURCES AND METHODS

The analysis is limited to spatial error of point locations using a simple error model. The consequence of spatial error on the outcome of overlay with datasets representing continuous variables is analysed. The suitability of the point data for overlay with a particular variable is assessed using the expected  $R^2$  value between the outcome of overlay given the spatial error, and the outcome of the overlay if the points were indeed accurately placed. This provides a measure of the *quality* of the point dataset *for a particular use*.

Quadrat locations (*source1*) were recorded in a floristic database (Turner and Muek 1992) in decimal degrees, calculated from map coordinates read from 1:100,000 scale topographic maps. *Source1* coordinate error is assumed to be due to incorrect transcription of points onto maps, incorrect reading of coordinates from maps, and incorrect transcription of coordinate values.

Quadrats locations (*source2*) were also marked onto 1:25,000 topographic maps of dubious cartographic lineage. It is assumed that the quadrat locations are marked correctly onto the topographic maps relative to local stream, ridge and road features identified on the map. *Source2* coordinate error is assumed to be due to stretches and shifts in the low quality map base.

*Source3* data were compiled by a separate process which avoids the sources of error associated with *source1* and *source2*. These are considered to be the true quadrat locations. Given the cartographic scale of the maps and the limits on the accuracy with which a point can be located on a map from a field survey, a planimetric error of 20 metres is regarded as very good.

Quadrat locations are illustrated in figure 1.

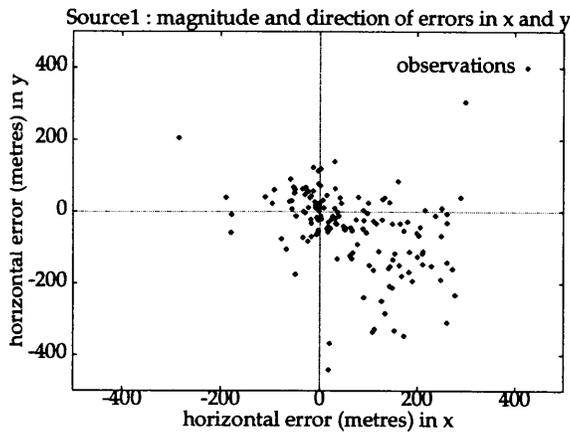


Figure 1. Location of study quadrats.

A terrain model (15 metre cell size) was interpolated from 1:25,000 contours (10 metre contour interval) for the study area using ANUDEM (Hutchinson 1988, 1989). Two attributes of the terrain model, slope and elevation, were estimated at each quadrat location using the GIS. Slope was estimated using the method of Burrough (1986) as implemented in ARC/INFO.

## INITIAL RESULTS

The spatial error associated with the points from *Source1* is illustrated in figure 2. Extreme values are not shown. *Source2* has similar errors.



**Figure 2.** Scatter diagram of the horizontal error in  $x$  and  $y$  observed in quadrat locations from *source1*. The diagram indicates that errors in  $x$  and  $y$  tend to be negatively correlated suggesting an anisotropic error distribution. The mechanisms behind this are unknown.

The importance of the spatial error in terms of inducing errors in the estimates of elevation and slope is illustrated by figure 3, which shows deviations from the correct values of slope and

elevation. Figure 3 clearly illustrates that for a given degree of spatial error the implications of the error are greater for estimates of slope than for elevation.

The results given in figure 3 can be summarised using  $R^2$  values for slope and elevation between the true locations and the mapped locations as follows:

	<i>Slope (percent rise)</i>	<i>Elevation (metres)</i>
<i>Point locations from Source1</i>	0.181	0.952
<i>Point locations from Source2</i>	0.318	0.976

## MODELLING THE IMPORTANCE OF SPATIAL ERRORS

Spatial error is present when  $\mathbf{x}_0 \neq \mathbf{x}_1$ ,  $\mathbf{x}_0$  being the true place of observation, and  $\mathbf{x}_1$  the location mapped. Let  $f(\mathbf{x})$  be a function denoting the probability density of the event  $\mathbf{x} = \mathbf{x}_1$  and  $Z(\mathbf{x}_1)$  be the actual observation (of slope, insolation, soil depth), while  $Z(\mathbf{x}_0)$  is the correct observation (unknown).

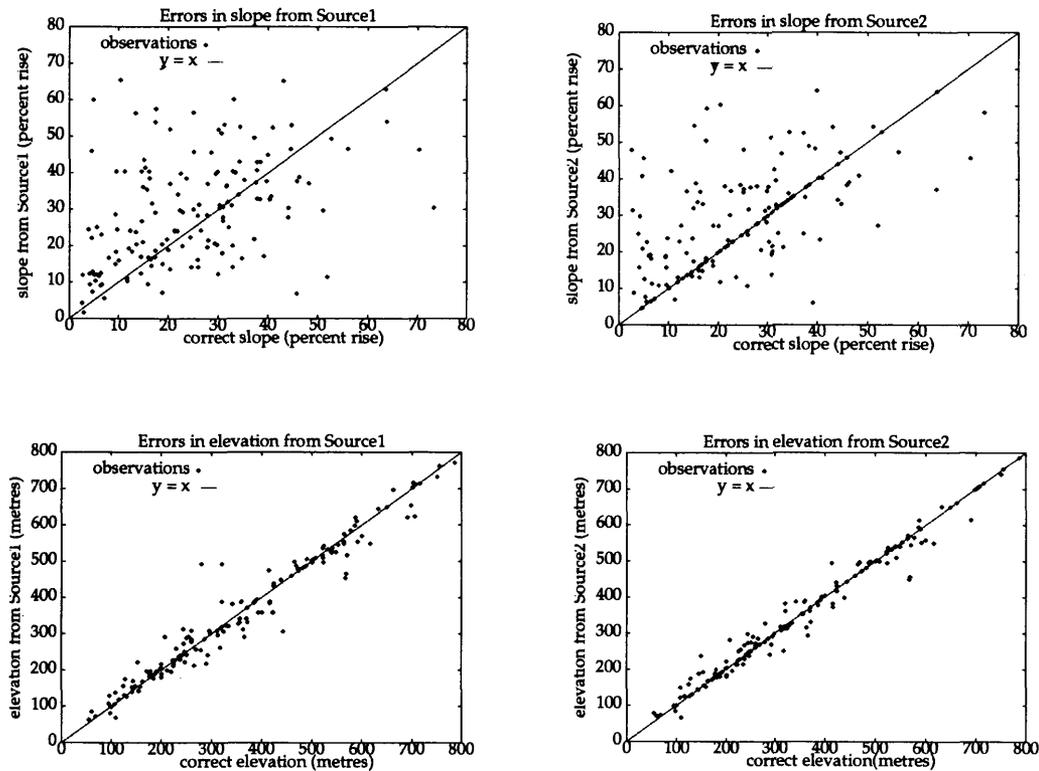
The question then, is: when  $\mathbf{x}_0$  is estimated by  $\mathbf{x}_1$ , what is the expected outcome of observations of some spatial variable,  $Z(\mathbf{x}_1)$ .

We can state the following:

$$E[Z(\mathbf{x}_1)] = \int Z(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = \int Z(\mathbf{x}_0 + \mathbf{h}) f(\mathbf{x}_0 + \mathbf{h}) d\mathbf{h} \quad (1)$$

where  $\mathbf{h}$  is a vector such that  $\mathbf{x} = \mathbf{x}_0 + \mathbf{h}$

This is the same form of equation identified by Allen and Starr (1982) as corresponding to observation of  $Z(\mathbf{x})$  at a coarser scale, with the scaling function weights being determined by  $f(\mathbf{x}_0 + \mathbf{h})$ . From this we can say that *spatial error in point locations has similar consequences for the expected value of observation as would viewing at a broader scale.*



**Figure 3.** Plots of estimates of elevation and slope from quadrat locations from *source1* and *source2*, against the correct estimate. Departures from 'y=x' indicate errors in the estimate due to quadrat location errors. Relatively large errors are introduced into estimates of slope *cf.* elevation, irrespective of the source of the quadrat error (*source1* vs. *source2*). The *type* of variable observed, rather than the specific amount of spatial error in the quadrats, appears to dictate whether the errors introduced will be substantial or negligible.

In qualitative terms, the results become fuzzy. Statistically, the correlation between the observed value  $Z(x_1)$  and the true value  $Z(x_0)$  will drop below 1, unless the spatial error is zero. The severity of the error will be indicated by the  $R^2$  value between the observed value,  $Z(x_1)$  and the true value  $Z(x_0)$ .

For a known spatial error,  $\mathbf{h}$  the covariance,  $cov(\mathbf{h})$ , between the imperfect observation, taken at location  $\mathbf{x}_1 = \mathbf{x}_0 + \mathbf{h}$ , and the true observation, taken at  $\mathbf{x}_0$ , is

$$cov(\mathbf{h}) = E[(Z(\mathbf{x}_0 + \mathbf{h}) - \mu_Z)(Z(\mathbf{x}_0) - \mu_Z)] \quad (2)$$

The correlation coefficient between  $Z(x_1)$  and  $Z(x_0)$  for a known error  $\mathbf{h}$  is given by

$$R(\mathbf{h})^2 = (cov(\mathbf{h}) / \sigma^2)^2 \quad (3)$$

where  $R(\mathbf{h})^2$  is commonly interpreted as the proportion of the variance in  $Z(x_0)$  accounted for by  $Z(x_1)$ .

The covariance is related to the semivariance by  $\gamma(\mathbf{h}) = \sigma^2 - cov(\mathbf{h})$ , (assuming that  $Z$  is a stationary function and that the error model is isotropic) so equation 3 can be re-stated as

$$[R(\mathbf{h})]^2 = (1 - \gamma(\mathbf{h}) / \sigma^2)^2 \quad (4)$$

To progress from the simple case of known error  $\mathbf{h}$  to a probability model of error  $f(\mathbf{h})$  we calculate the expected value of  $cov(\mathbf{h})$  using simply:

$$\begin{aligned} E[R^2] &= \int f(\mathbf{h}) R(\mathbf{h})^2 d\mathbf{h} \\ &= \int f(\mathbf{h}) (1 - \gamma(\mathbf{h}) / \sigma^2)^2 d\mathbf{h} \end{aligned} \quad (5)$$

This gives a direct mathematical expression to estimate the importance of spatial error in a given situation, requiring only a model of the spatial error of the point observations and a correlogram or variogram of the spatial variable  $Z$ . The semivariance,  $\gamma(\mathbf{h})$  can be thought of as that part of the total variance which is introduced at lag  $\mathbf{h}$ .

Equation 5 also illustrates the relationship between the magnitude of the spatial error and the inherent scaling of the variable in determining the importance of the spatial error, discussed earlier. Clearly if  $f(\mathbf{h})$  is high only where  $\gamma(\mathbf{h})$  is near 0, the error is less important than if  $f(\mathbf{h})$  is high over a wide range of  $\gamma(\mathbf{h})$  including where  $\gamma(\mathbf{h}) \sim \sigma^2$ . Forms of  $f(\mathbf{h})$  which either have a non-zero mean error,  $E[\mathbf{h}] \neq \mathbf{0}$ , or which are multi-modal, with a peak some distance from  $\mathbf{x}_0$  (eg. Bolstad *et al.* 1990) will tend to introduce more significant errors.

### **Prediction of the errors in elevation and slope introduced through spatial error in coordinate locations.**

Equation 5 demonstrates that information on the distribution of spatial errors in the point data and a knowledge of the variogram of  $Z$  are required to estimate  $R^2$ , giving a direct assessment of the *data quality* of the points, for observation of a given variable  $Z$ .

If isotropic variograms and error models are assumed, equation 5 simplifies to

$$E[R^2] = \int f(h) R(h)^2 dh = \int f(h) (1 - \gamma(h) / \sigma^2)^2 dh \quad (6)$$

where  $h = |\mathbf{h}|$ .

In the following section this method is applied to the datasets used here. The Excel software was used to manipulate data and to apply mathematical formulae, while variograms were estimated using software written by the author. The continuous functions indicated by equation 6 were approximated using discrete forms with an interval of 40 metres ( $h=0,40, \dots,8000$ ).

### **Models of spatial error**

An isometric error model was adopted. Normal and lognormal distributions were fitted to observed errors using the method of moments (figure 4).

The parametric models enable the standard deviation to be used as a parameter to vary the model, allowing sensitivity testing. The assumption of an isometric normal model,  $f(h)$ , has the further advantage of requiring only one parameter,  $\sigma_e$ . The convention of  $epsilon = 2\sigma_e$  is adopted, thus in terms of an

*epsilon* model of error we can state that 95% of the data points are expected to lie within *epsilon* metres of the mapped point.

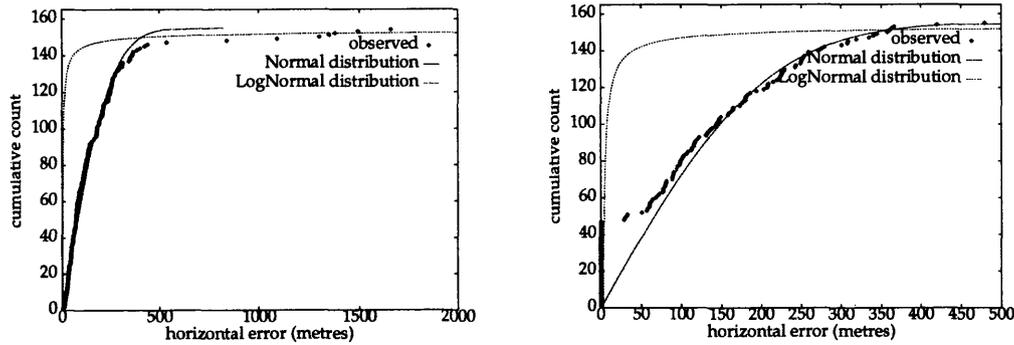


Figure 4. Plots of the cumulative distribution of horizontal error for *source1* quadrats (left) and *source2* quadrats (right); and normal and lognormal probability models fitted using the method of moments. An isotropic error distribution is assumed. For *source1*, the normal distribution was fitted after exclusion of extreme values. Neither probability model deals with *source1* very well, however the normal distribution was chosen in preference to the lognormal.

#### ASSESSING DATA QUALITY FOR SLOPE AND ELEVATION

The discrete version of equation 6 was used to calculate the expected  $R^2$  value for true and estimated values of slope and elevation from *source1* and *source2* for the values of *epsilon* given in Table 1.

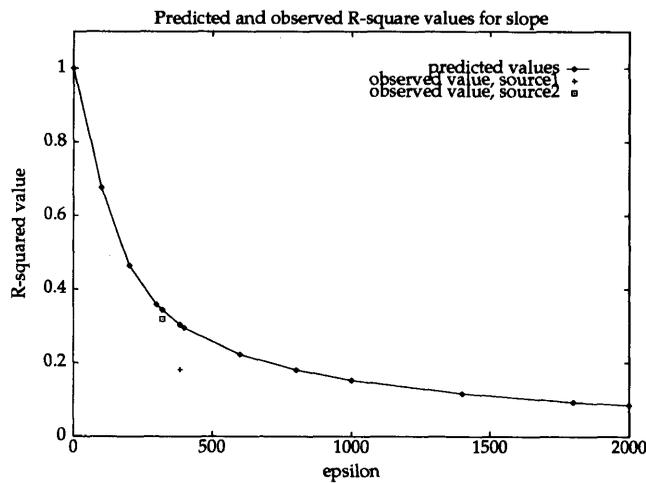
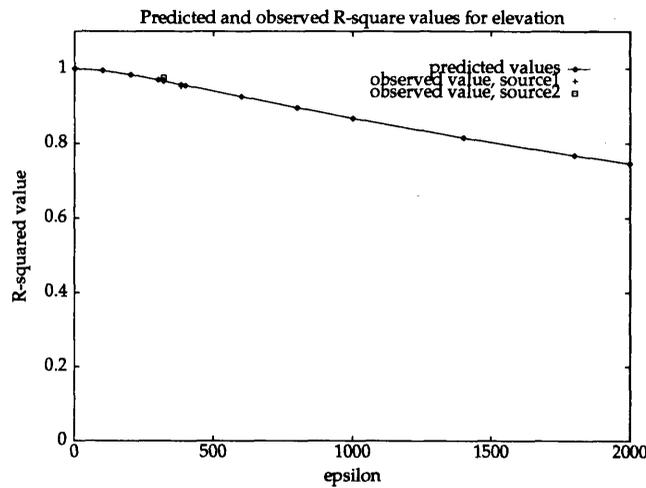
$$E[R^2] = \sum_{h=\Delta, 2\Delta, \dots, 8000} [F(h+\Delta/2) - F(h-\Delta/2)][1 - \gamma(h) / \sigma^2]^2 + [F(\Delta/2) - F(0)][1 - \gamma(0) / \sigma^2]^2 \quad (7)$$

where:  $F(h)$  is the cumulative normal probability distribution with standard deviation of  $\epsilon/2$ , truncated to exclude negative values ( $F(0) = 0$ ) and scaled to maintain  $F(\infty) = 1$ ;  $h$  is the magnitude of the spatial error, and the lag distance for  $\gamma(h)$ ;  $\Delta$  is an interval (40 m),  $\sigma^2$  is the observed variance of  $Z$  (slope or elevation, as appropriate).

Using equation 7 the predicted  $R^2$  values for the observations of slope and elevation, from *source1* and *source2*, were calculated and compared with the observed values (table 1).

Table 1. Modelled and observed (in brackets) correlations ( $R^2$ ) between observations of elevation and slope from true data point locations and data point locations containing spatial error.  $\epsilon = 2\sigma$ .

	Slope (percent rise)	Elevation (metres)
Point locations from <i>Source1</i> . <i>Epsilon</i> = 384	0.303 (0.181)	0.957 (0.952)
Point locations from <i>Source2</i> . <i>Epsilon</i> = 320	0.343 (0.318)	0.967 (0.976)



In most practical situations,  $\gamma(h)$  is readily estimated for datasets in a GIS, but the precise nature of  $F(h)$  will be unknown. To develop some idea of the sensitivity of data quality to  $f(h)$ ,  $E[R^2]$  was calculated for a wide range of  $\epsilon$ . Results are shown in figure 5, from which it is clear that the modelled  $R^2$  values for elevation from *source1* and *source2* are closely predicted.  $R^2$  for slope is accurately predicted for *source2*, but for *source1* is optimistic, reflecting the extreme values not catered for by the normal distribution (figure 4).

**Figure 5. Predictions of  $R^2$  value for a wide range of error bands, compared with observed  $R^2$  values.**

The contrast between figures 5 and 6 shows that the key

factor determining the  $R^2$  between correct and incorrect values of elevation and slope is *not* the spatial error in the observation points, but *the inherent spatial structure of the variables being observed*.

Thus, even the very coarse models of spatial error used here are sufficient to quite accurately model the quality of the quadrat data sources for overlay with other spatial data themes.

## DISCUSSION

These results have wide-ranging practical value. They suggest that the detail of the probability error model  $f(x)$ , is less important in determining data quality than the inherent properties of the surfaces being interrogated. Only the main features of  $f(x)$  seem to be important.

The results also suggest an approach to management of GIS error which integrates spatial and attribute error. If the value of a variable is estimated from interrogation of a data-surface  $Z$ , where  $Z$  is a model with residual (attribute) error  $\sigma^2_{eZ}$ , the variance of  $Z$ , at any given place, due to spatial and attribute error can

be calculated as the expected semivariance, plus the variance of the residual error of the model.

$$\sigma^2_Z = E[\gamma_Z] + \sigma^2_{e_Z} \quad \text{where} \quad E[\gamma_Z] = \int f_Z(\mathbf{h}) \gamma_Z(\mathbf{h}) d\mathbf{h} \quad (8)$$

In which  $f_Z(\mathbf{h})$  is the pdf of the spatial errors in variable Z, while  $\gamma_Z(\mathbf{h})$  is the semivariance of variable Z readily estimated readily from the data-surface Z.

Applying equation 8 to the DEM used here, assuming isometric normally distributed spatial errors with *epsilon* = 25, and an RMS of 5 ( $\sigma^2_{e_Z} = 25$ ) gives  $E[\gamma_Z] \approx 3.7$ ,  $\sigma^2_Z = 3.7 + 25$ . Thus (for elevation) the influence of spatial error is small compared with residual error.

## REFERENCES

- Allen T F H, Starr, T B (1982) *Hierarchy: perspectives for ecological complexity*. University of Chicago Press
- Bolstad, P V, Gessler, P, Lillesand, T M (1990) Positional uncertainty in manually digitised map data. *International Journal of Geographic Information Systems*. 4 399
- Burrough, P A. (1986) *Principles of Geographical information systems for land resources assessment*. Clarendon Press. Oxford.
- Goodchild, M F. The issue of accuracy in global databases. (1988) In Mounsey, H, Tomlinson, R F. (eds.) *Building databases for global science. Proceedings of the first meeting of the International Geographical Union Global Database Planning Project*. Tylney Hall, Hampshire, UK, May 1988. Taylor and Francis. London & New York, 1988
- Goodchild, M F, Guoqing, S, Shiren, Y. (1992) Development and test of an error model for categorical data. *International Journal of Geographical Information Systems*. 6 87-104
- Hutchinson, M F (1988) Calculation of hydrologically sound digital elevation models. *Proceedings of the Third International Symposium on Spatial Data Handling*. Sydney, Australia.
- Hutchinson, M F (1989) A new procedure for gridding elevation and stream line data with automatic removal of spurious pits. *Journal of Hydrology* 106 211-232
- Turner, L A, Mueck, S G. (1992) *The vegetation of the Sardine, Rich and Ellery forest blocks, Orbost region, Victoria*. Silvicultural Systems Project Technical Report number 9. Department of Conservation and Environment, Melbourne, Australia.

## BIOGRAPHICAL SKETCH

Adam Lewis is the Senior GIS Scientist in the Natural Resource Systems Branch (NRS) of the Department of Conservation and Natural Resources, based in Melbourne, Australia. He has experience in Forest and Land Management, and recently completed a PhD and the Australian National University.

Michael Hutchinson is a Senior Fellow at the Australian National University whose primary interest is the spatial and temporal analysis of physical environmental data.

# Moving Into Secondary Map Projections - An Analysis of Potential Inconsistencies In Spatial Data

Mohamad N. Said<sup>1</sup> and Peter F. Fisher<sup>2</sup>

---

**Abstract.**—Implementation of GIS functions especially those involving complex mathematical computation such as map projection may result in a certain amount of deformation in spatial data. In this research, several experiments using a test pattern are carried out to test the map projection function provided by two selected commercial GIS's. Experiments include the change of projection within the same and across categories. It is found that there are geometrical and topological inconsistencies as a results of these transformations. Most are relatively small and probably the result of digital rounding errors through the computational process.

---

## INTRODUCTION

All spatial data layers stored in a Geographical Information Systems (GIS) are based on a certain type of map projection which defines an x-y Cartesian coordinate system for all features contained in the database. Three categories of map projection are commonly identified: conformal projections maintain the shape; equal area projections maintain the area; and equidistant projections preserve the distance.

Having a multiple-theme database, therefore, requires all themes to be in a common projection, and if they are not some number will need to be transformed to fit a selected one. Such transformations are common in practice and may take place a number of times, especially when GIS users exchange digital map data or require different properties in their output maps. The reprojection may involve a change of parameters contained in the output projection including the geodetic datum, the central meridian, the standard parallels, the origin etc.

---

<sup>1</sup> Research student, Dept. of Geography, University of Leicester, U.K.

<sup>2</sup> Senior lecturer, Dept. of Geography, University of Leicester, U.K.

It is anticipated that the reprojection of original maps into secondary projections may cause a certain amount of deformation and the magnitude and pattern of the errors may be different when transformations are done using different GIS packages. This paper reports the outcome of the research being carried out to demonstrate these potential inconsistencies and discrepancies.

## RELATED WORKS

Some works on benchmarking of GIS functions or algorithms do exist in the literature. Piwowar et al. (1990) reported the research outcome of raster-to-vector conversion using three different algorithms while Knaap (1992) took further tests of the same work by comparing the results output by eight different GIS packages. Knaap showed that no two of the output produced the same rasterization of the test vector patterns, although there were some reassuring consistencies. Fisher (1993) explored the viewshed functions and reported the extent of variability in the viewable area when the function is implemented in different packages. He further highlighted the need for standards and/or empirical benchmark datasets for GIS functions and the desirability of publications of algorithms used in GIS functions.

Franklin (1984) demonstrated the topological effects of changing scale and rotation when rounding occurs in integer arithmetic. Burrough (1986) quoted the work by Gruenberger (1984) of how computer word length can have a cumulative effect after a real number is repeatedly squared on different platforms.

Masters (1992) stressed the importance of spatial data accuracy which has to be considered in the context of coordinate systems defined by geodetic datums and map projections. He further made a comment on the continuing difficulties in GIS projects due to the assumption made by many GIS's that the coordinate systems that maps are based on are all the same and all have the same accuracy.

## EXPERIMENTAL DESIGN AND TEST DATA

### The Map Projections

In the implementations tested in this study, three projections are selected. These are Lambert Conformal Conic (or Lambert), Universal Transverse Mercator (or UTM) and Albers Conic Equal Area (or Albers) projections. The parameters used throughout the experiments are as follows:

#### Lambert Conformal Conic

Spheroid	Clarke 1866
Units	meter
Central Meridian	75 <sup>0</sup> 00" 00' W

Standard Parallels	2 <sup>0</sup> 20' 00"N
	3 <sup>0</sup> 40' 00"N
Latitude of Projection Origin	0 <sup>0</sup> 00' 00"
False Easting (m)	0.0000
False Northing (m)	0.0000

Universal Transverse Mercator (UTM)

Spheroid	Clarke 1866
Units	meter
Zone	18

Albers Equal Area

Spheroid	Clarke 1866
Units	meter
Central Meridian	75 <sup>0</sup> 00' 00" W
Standard Parallels	2 <sup>0</sup> 20' 00"N
	3 <sup>0</sup> 40' 00"N
Latitude of Projection Origin	0 <sup>0</sup> 00' 00"
False Easting (m)	0.0000
False Northing (m)	0.0000

**The Test Data**

An experimental test pattern of 16 square polygons (figure 1) covering 2<sup>0</sup> by 2<sup>0</sup> in the equatorial region is designed to fit the Zone 18 band of the United States (longitudes: 74<sup>0</sup>W to 76<sup>0</sup>W). The lower limit of the latitude is at 2<sup>0</sup> 00' 00"N while the upper limit is at 4<sup>0</sup> 00' 00"N. The Central Meridian of this Zone is chosen to be at the longitude of 75<sup>0</sup>W which is along the central line of the quadrangle.

**The GIS Packages**

Two commercial GIS packages are used in the study, namely Arc/Info (version 7.0.3) and Genamap (version 5.1). Arc/Info runs on an IRIX mainframe while Genamap does on an IBM RISC6000 workstation. The original test data are however created and keyed independently to avoid any deformation in the data transfer process.

**The Transformation Schemes**

Experiments are carried out to test for any inconsistencies in the conformal map (Lambert) when it is transformed into other projections within its category as well

as across the other categories. The transformations are accomplished according to the following schemes:

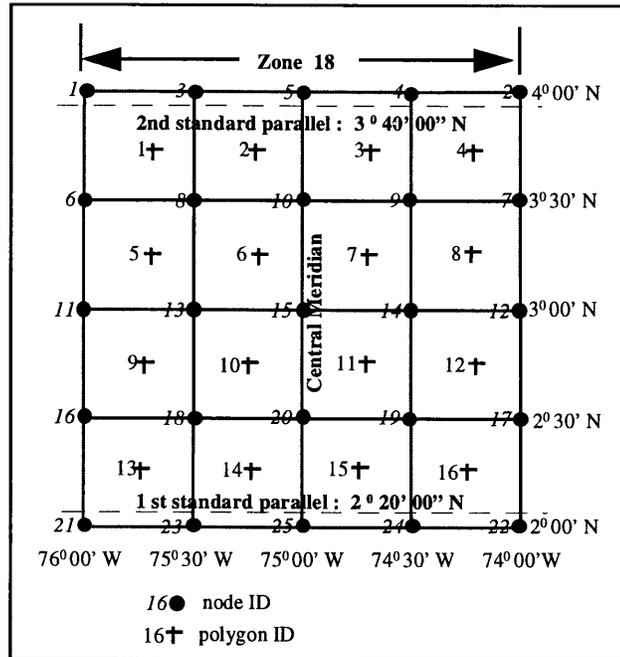


Figure 1. -- The test pattern

a) Transformation of geographical coordinates into Lambert Conformal Conic

**GEOG** ----> **LMB1**

GEOG represents the original map in geographical coordinates (latitude and longitude) which is transformed into Lambert Conformal Conic projection producing **LMB1**.

b) Transformation Within Conformal Category

**LMB1** ----> **UTM** ----> **LMB2**

The initial map in Lambert projection as obtained in (a) is further transformed into another conformal projection, in this test **UTM** is chosen. This map is then reprojected back into Lambert Conformal Conic producing **LMB2**.

c) Transformation Across Category

**LMB1** ----> **ALBERS** ----> **LMB3**

In this case the interchange projection is equal area (*ALBERS*). This produces the third Lambert map named as **LMB3**.

The purpose of reprojection as carried out in (b) and (c) above is to enable checking of inconsistency between **LMB1** and **LMB2**.

Further experiments are also made by overlaying maps of similar projection before and after transforming. This includes a union process of combining **LMB1<sub>A</sub>** and **LMB1<sub>G</sub>** as well as **LMB2<sub>A</sub>** and **LMB2<sub>G</sub>** (the subscripts A and G denote maps transformed with Arc/Info and Genamap GIS's respectively). Any new intersection points will indicate the occurrence of topological errors.

**Table 1.—Internal consistency check - shifts in coordinates.**

Node ID	Arc/Info		Genamap	
	LMB2 - LMB1		LMB2 - LMB1	
	diff in x (m)	diff in y (m)	diff in x (m)	diff in y (m)
1	-0.0070	0.0000	0.0000	0.0000
2	0.0070	0.0000	0.0000	0.0000
3	-0.0150	0.0000	0.0000	0.0000
4	0.0150	0.0000	0.0000	0.0000
5	0.0000	0.0000	0.0000	0.0000
6	0.0000	0.0000	0.0000	0.0000
7	-0.0320	0.0000	0.0000	0.0000
8	-0.0040	0.0000	0.0000	0.0000
9	-0.0280	0.0000	0.0000	0.0000
10	0.0000	0.0000	0.0000	0.0000
11	0.0080	0.0000	0.0000	0.0000
12	0.0230	0.0000	0.0000	0.0000
13	-0.0040	0.0000	0.0000	0.0000
14	0.0040	0.0000	0.0000	0.0000
15	0.0000	0.0000	0.0000	0.0000
16	-0.0150	0.0000	0.0000	0.0000
17	-0.0160	0.0000	0.0000	0.0000
18	0.0080	0.0000	0.0000	0.0000
19	-0.0080	0.0000	0.0000	0.0000
20	0.0000	0.0000	0.0000	0.0000
21	-0.0150	0.0000	0.0000	0.0000
22	-0.0160	0.0000	0.0000	0.0000
23	0.0080	0.0000	0.0000	0.0000
24	-0.0080	0.0000	0.0000	0.0000
25	0.0000	0.0000	0.0000	0.0000

## RESULTS

Analyses are carried out on the resulting Lambert maps, represented in bold in previous section. Observations are made on any shifts in the coordinates of nodes as well as in the area of polygons contained in the maps.

The internal consistency of each software is checked by comparing the first and the second Lambert maps (see transformation schemes (b) and (c)). From table 1, it can be seen that transformation between conformal projections with Arc/Info produces some amount of error in the feature coordinates. These errors occur only in the x direction and are very small, with a maximum magnitude of 0.032m, relative to the coverage of the study pattern. As a consequence of these shifts, the polygons suffer a change in area between 256 to 1280 square meters. Genamap's output seems to be more consistent for this particular test, where no errors are shown. Transformation via the equal area projection however does not cause any error in either package.

**Table 2.—Comparison of absolute coordinates: Arc/Info vs Genamap**

Node	LMB1 <sub>G</sub> - LMB1 <sub>A</sub> (m)		LMB2 <sub>G</sub> - LMB2 <sub>A</sub> (m)		LMB3 <sub>G</sub> - LMB3 <sub>A</sub> (m)	
	Diff. in x	Diff. in y	Diff. in x	Diff. in y	Diff. in x	Diff. in y
1	-0.0011	0.0128	0.0059	0.0128	-0.0011	0.0128
2	0.0011	0.0128	-0.0059	0.0128	0.0011	0.0128
3	0.0017	0.0014	0.0167	0.0014	0.0017	0.0014
4	-0.0017	0.0014	-0.0167	0.0014	-0.0017	0.0014
5	0.0000	0.0073	0.0000	0.0073	0.0000	0.0073
6	-0.0035	-0.0110	-0.0035	-0.0110	-0.0035	-0.0110
7	0.0035	-0.0110	0.0355	-0.0110	0.0035	-0.0110
8	0.0005	-0.0086	0.0045	-0.0086	0.0005	-0.0086
9	-0.0005	-0.0086	0.0275	-0.0086	-0.0005	-0.0086
10	0.0000	-0.0085	0.0000	-0.0085	0.0000	-0.0085
11	-0.0015	-0.0053	-0.0095	-0.0053	-0.0015	-0.0053
12	0.0015	-0.0053	-0.0215	-0.0053	0.0015	-0.0053
13	0.0015	0.0107	0.0055	0.0107	0.0015	0.0107
14	-0.0015	0.0107	-0.0055	0.0107	-0.0015	0.0107
15	0.0000	0.0061	0.0000	0.0061	0.0000	0.0061
16	0.0010	0.0120	0.0160	0.0120	0.0010	0.0120
17	-0.0010	0.0120	0.0150	0.0120	-0.0010	0.0120
18	-0.0013	0.0107	-0.0093	0.0107	-0.0013	0.0107
19	0.0013	0.0107	0.0093	0.0107	0.0013	0.0107
20	0.0000	0.0003	0.0000	0.0003	0.0000	0.0003
21	0.0000	-0.0062	0.0150	-0.0062	0.0000	-0.0062
22	0.0000	-0.0062	0.0160	-0.0062	0.0000	-0.0062
23	0.0022	0.0063	-0.0058	0.0063	0.0022	0.0063
24	-0.0022	0.0063	0.0058	0.0063	-0.0022	0.0063
25	0.0000	0.0051	0.0000	0.0051	0.0000	0.0051

A comparison of absolute coordinates is shown in table 2. There is a certain amount of discrepancy in both x and y coordinates. The pattern of these differences, in all three comparisons, is found to be symmetrical about the central meridian. These discrepancies have caused the change in area of the polygons between 10 to 1024 square meters. The small differences seen above are sufficient to cause line intersections when maps from different packages are combined. For example, by combining **LMB1<sub>A</sub>** and **LMB1<sub>G</sub>**, there are 68 polygons created instead of 16. This means 52 sliver polygons have been generated even though they are small, i.e. with a maximum area of about 638 square meters. It is worse when **LMB2<sub>A</sub>** and **LMB2<sub>G</sub>** are combined, where 72 polygons are generated giving an extra of 56 with a maximum area of about 939 square meters. Graphically, these polygons are invisible without considerable magnification but each of them adds up one new records in the attribute database. Thus in the first example, 52 new records occur in the attribute database.

## CONCLUSIONS

From the various tests described and implemented throughout this research, it has been shown that errors do exist as a results of moving from one projection to another. They are found to be very small and may be due to digital rounding. It is believed that different GIS packages may or may not introduce internal inconsistencies but they may produce different results when compared to each other. Further experiments such as transformations of maps at varying latitudes and with other projection combinations have been examined.

Integrating datasets using derived maps, in this case maps with secondary projections, from different packages may cause spatial problems such as line intersections and sliver polygons. This problem is inherent in the maintenance of the spatial and attribute databases when new unwanted features are generated. It has been shown that with small discrepancies in feature coordinates, the database has been grown by three times.

As a final conclusion, further benchmarking tests on map projection function with other GIS packages will add to the knowledge of potential inconsistencies in the spatial data. This will reassure the GIS users about the function they are using.

## REFERENCES

- Briggs, D. and Mounsey, H., 1989. Integrating Land Resource Data Into A European Geographical Information System : Practicalities and Problems. Applied Geography, Vol. 9 (1989). pp 5-20.
- Burrough, P. A. 1986. Principles of Geographical Information Systems For Land Resources Assessment. Clarendon Press: Oxford. 194p.

- Fisher, P. F. 1993. Algorithm and implementation Uncertainty In Viewshed Analysis. International Journal of Geographical Information Systems, Vol. 7, No. 4. London: Taylor and Francis. pp 331-347.
- Franklin, R., 1984. Cartographic Errors Symptomatic of Underlying Algebra Problems - Proceedings of the International Symposium On Spatial Data Handling, 1994, Zurich. pp 190-208.
- Knaap, W.G. M. VAN DER, 1992. The Vector To Raster Conversion: (mis)use In Geographical Information Systems. International Journal of Geographical Information Systems, Vol. 6, No. 2. London: Taylor and Francis. pp 159-170.
- Maling, D.H., 1991. Coordinate Systems And Map Projections For GIS - in Geographical Information Systems: Principles And Applications, Vol. 1 (ed. Maguire, D.J., Goodchild, M.F. and Rhind, D.W). Longman Group UK Ltd.
- Masters, M., 1992. Digitization and Database Accuracy. Proceedings of GIS/LIS '92 Annual Conference and Exposition, San Jose, California. pp 522-531.
- Piwowar, J.M., LeDrew, E.F., and Dudycha, D.J. , 1990. Integration of Spatial Data In Vector and Raster Formats In A Geographic Information System Environment. International Journal of Geographical Information Systems, Vol. 4, No. 4. London: Taylor and Francis. pp 429-444.
- Snyder, J.P. 1987 Map Projections - A Working Manual. U.S. Geological Survey Professional Paper 1395. p383.

### **BIOGRAPHICAL SKETCH**

Mohamad N. Said is currently a research student, doing his PhD in GIS with particular interest in studying the positional errors in spatial data due to the change of map projection systems. He obtained his bachelor degree in Surveying and Mapping Science from the North East London Polytechnic, U.K. (1985) and master degree from Curtin University of Technology, Australia (1990).

Peter F. Fisher is a senior lecturer at the Department of Geography, University of Leicester with special interests in error modelling, propagation and visualization. He is currently the editor of International Journal of Geographic Information Systems. He holds a B.Sc. in Environmental Sciences from the Lancaster University (1977), an M.Sc. in Pedology and Soil Survey from Reading University and a PhD from Kingston Polytechnic (1982).

# Filling in Missing Forestry Data: Exploring Autocorrelational Techniques

Alissa N. Antle<sup>1</sup> and Peter L. Marshall<sup>2</sup>

---

**Abstract.**—A crucial component of any effective approach to long-term management and protection of valuable forest resources is the ability to represent and model spatial forest ecosystem processes. Geographic information systems (GIS) integrated with spatial forest models provide an appropriate spatial framework for forest ecosystem modeling. In order to be effective, accurate ecological data must be used as input to GIS and spatial forest models. It is often the case that the forest inventory attribute of interest may have incomplete coverage for the area under management. In these cases it is mandatory to have techniques that can predict the missing values while preserving the complex spatial patterns between forest inventory attributes. One of the most prominent spatial patterns found in ecological data is the presence autocorrelation. This paper introduces several autocorrelational tests, describes different methods of representing spatial autocorrelational structures, and discusses various approaches which, either explicitly or implicitly, use this information to predict missing values in ecological data sets.

---

## INTRODUCTION

In order for spatial forest models to be effective, accurate and complete ecological data must be used. Spatial data quality has received much recent attention in the GIS community (Goodchild and Gopal 1989, Chrisman 1991, Guptill and Morrison 1995). The term completeness has been used in reference to both resolution and missing values. Missing values are those which are expected to appear in the data set but do not. In forest inventory data sets, values may be missing due to collection errors or omissions. In particular, attributes required for a specific forest model may have not been collected, or because of the expense or effort required to collect the data, only a limited set of data may be available. In order to ensure that data sets are fit for forest modeling purposes, techniques that can predict missing values are necessary to complete data sets prior to use in

---

<sup>1</sup> *Department of Geography, University of British Columbia, Canada*

<sup>2</sup> *Department of Forest Resources Management, University of British Columbia, Canada*

existing models or in developing new models.

While conventional statistical techniques have been widely used to predict missing values, most of these techniques substitute mean or median values for missing data. These approaches do not preserve the complex spatial patterns which may be found among multivariate forest inventory attributes. In addition, most interpolative approaches are based on the assumption of spatially independent data. Simply stated, the problem is how to best predict missing, autocorrelated data.

This paper presents an initial exploration into the missing forestry data problem given the autocorrelated nature of most spatial ecological data. Space restrictions preclude a complete discussion of this topic.

## **SPATIAL STRUCTURE**

While the problem outlined above may seem to be a simple statistical inference problem, it is not. The distribution of natural abiotic, and induced biotic features, are typically neither uniform nor random. Rather, they tend to be either aggregated in patches or form gradients or other types of spatial structures (Legendre 1993).

Why is representation of spatial pattern important in forest and ecological models? The spatial or temporal structure or pattern of ecosystems is a crucial element in most ecological theories. Any forest or ecological model based on such theories must take spatial pattern and the nature of spatial ecological data into account. Specific characteristics of spatial data can be identified which cause problems in traditional statistical techniques. There are two main classes of spatial effects: spatial dependence and spatial heterogeneity (Anselin and Getis 1992, Cressie and ver Hoef 1993). Only spatial dependence will be considered in this paper.

### **Spatial Autocorrelation**

What is significant about the patterns found in spatial ecological data? According to Tobler's first law of geography, everything is related to everything else, but near things are more related than distant things. This simply means that the relations within and between variables sampled at spatially proximal locations will be stronger than those sampled at further distances. Ecological data are spatially dependent by nature. Spatial dependence is often referred to as spatial autocorrelation.

A major consequence of spatial autocorrelation is that traditional statistical inference techniques are not as efficient as an independent sample of the same size. That is, spatial dependence leads to a loss of information, information which is crucial for predicting missing values. Positive spatial autocorrelation may cause the variance to be underestimated. Negative autocorrelation may inflate estimates of variance. In either case, predictive equations can become biased and the results

of prediction unreliable.

Clearly, some new methods or enhancements to existing statistical methods of spatial inference are needed to address the case of spatially autocorrelated data (Antle and Klinkenberg 1995).

## MODELING SPATIAL AUTOCORRELATION

As discussed above, many of the basic statistical methods used in ecological studies are impaired by autocorrelated data. Most, if not all, environmental data falls into this category (Legendre and Fortin 1989). Prediction methods which can be used to fill in missing forestry data must account for spatial autocorrelation or be resistant to its effects. Before these techniques are applied, the nature of the spatial patterns in the data should be explored. Various tests for autocorrelation and trends can be used. Once we know that the data is correlated and the nature of the relation (i.e., either positive or negative), techniques which take this pattern into account can be applied.

### Testing for Autocorrelation

Since the presence of autocorrelation may distort the results of conventional statistical interpolation techniques, autocorrelation should be tested for prior to selecting any particular model (i.e., during the exploratory phase of data analysis) in order to ensure accurate predictions. There are various methods which test for the presence of spatial autocorrelation in univariate data and for correlation in multivariate data sets. These are well documented and will not be described here. Univariate methods include the joint count test for nominal data, Geary's  $c$  or Moran's  $I$  test for ordinal and interval data, and spectral analysis (see Haining 1990 for a detailed description). Multivariate methods include the Spearman rank correlation coefficient for ordinal data, the Pearson product correlation coefficient for interval data, the Mantel test and correlogram (see Haining 1990 or Goodchild 1986 for details).

If large scale spatial dependency exists, it may be removed by regression or model-fitting (eg., trend surface analysis). Traditional statistical techniques can then be used without violating their assumptions. However, caution should be taken in order to avoid removing the determinants of the underlying processes along with the spatial structure. An alternative to removing spatial dependence is to modify statistical approaches to take spatial autocorrelation into account. At the very least, techniques which are resistant to spatial autocorrelational effects should be used. Techniques which utilize autocorrelational information (to enhance prediction accuracy) are preferable. First how do we represent spatial autocorrelation?

## **Representing Autocorrelation**

Autocorrelation can be represented through structure functions, which help quantify spatial dependency and partition it along distance intervals.

### **Geographic coordinates**

The spatial structure of variables can be expressed as a linear combination of geographic coordinates of sample sites (Legendre 1993). For example, a high order polynomial can be built up using the  $x$  and  $y$  coordinates of sites (as is done in trend surface analysis). This information can then be incorporated into traditional inference models, such as regression models, using partial regression analysis techniques. The result of this analysis is the separation of the variation of the target variable into four components: the variance resulting from non-spatial environmental factors, spatial structure, the interaction of environmental and spatial factors, and the unexplained variation.

### **Geographic distances**

Another approach is to use location as one of a set of predictor variables in a statistical model (Legendre 1993). In this case, the spatial structure of the data is represented by a matrix of geographic distances between site locations (also called a proximity matrix). For example, the euclidean distance can be computed for all pairs of sites based on their geographic coordinates. This information is then summarized in a spatial distance matrix. If the remaining environmental variables can be represented in the form of distance matrices, then these distance matrices can be compared using some form of correlation analysis (e.g., the Mantel test, see Legendre and Fortin 1989).

### **Covariation**

One of the most common structure functions is the variogram (and related covariance and correlogram). The variogram summarizes the spatial continuity for all possible pairings of data, for all significant lag distances, by modeling the average degree of similarity between values as a function of their separation distance. Variograms can be computed either as averages over all directions or specific to a particular direction. If there is a trend in the data (i.e., local means and local variances change as a function of location within the sampling space), the variogram will include both lag-to-lag variability and the trend variability (Rossi *et al.* 1992). In the case of stationary data, the variogram can be experimentally estimated from observed values. Various statistical models can then be fitted to the variogram, including exponential, spherical and linear.

## **Prediction Utilizing Autocorrelation**

Since ecological data are often correlated, techniques which can take

advantage of the information contained in these relationships may be able to best predict missing values while retaining the variability and covariation structure of forest attribute data.

### **Regression models**

Traditional linear regression can be extended to situations where spatial autocorrelation exists by the addition of lagged variables. Griffith (1987) outlines how spatial autocorrelation can be incorporated into traditional regression models. Promising results have been obtained from the introduction of an additional parameter into conventional statistical models that accounts for the latent spatial autocorrelative structure of the data. This technique has also been extended to multivariate models (Haining 1990). This approach is one step toward the integration of spatial dependence (in the form of geographic distance information) into predictive models.

As mentioned above, in partial regression analysis, the spatial component can be "partialled out" by regressing the spatial factor variables onto each explanatory variable until only the regression residuals remain (Legendre 1993). The residuals are then used to model the variable(s) which have missing values. This approach "explains" much of the variation that is left unaccounted for in simple regression models.

### **Kriging**

Kriging is a well known geostatistical spatial interpolation technique which explicitly uses autocorrelational information in its predictions (Rossi *et al.* 1994). Since it uses a local estimator and the autocorrelation structure of the data, kriging is quite unlike traditional methods, such as trend surface analysis, which ignore autocorrelational effects and use all of the data to estimate the unknown point's value. In addition, kriging provides variance information which is crucial in determining the accuracy of predicted values.

Kriging can be applied to both nominal and continuous data. In general, kriging predicts missing values at a specific site by taking a weighted linear average of available samples (like regression). It is similar to multiple regression with a few important differences. First, kriging can produce predicted values that are either larger or smaller than any of the sample values. Second, kriging utilizes both distance and geometry among samples, whereas traditional methods use only distance. Third, kriging attempts to minimize the variance of the expected error by inferring the variance from an empirical model of degree of spatial dependence with distance and direction (i.e., the variogram).

### **Most similar neighbor analysis**

Most similar neighbor (MSN) analysis is a canonically based technique which can be used to fill in interval or continuous data values by implicitly utilizing the spatial structural information of the data. Instead of estimating missing values one-

by-one (for a particular site), the technique "chooses" a most similar site from a set of sites to act as a stand-in. The surrogate site is chosen on the basis of a similarity function (derived using canonical correlation analysis).

This approach works well when the variable to be predicted and the predictor variables have strong functional relationships to each other (Moeur 1995) (as is the case for spatially autocorrelated data). Because surrogate sites are chosen from among actual sites, impossible predictions for missing values cannot occur. However, this means that the samples sites must adequately span the full range of sites. While Moeur's procedure does not explicitly model spatial correlation, it attempts to maintain the covariance structure of the multivariate response attributes.

## DISCUSSION

None of the methods described here for filling in missing forestry data can be used without careful consideration. They all make assumptions about the nature of data and the phenomena under study. For specific cases, these assumptions may or may not hold. The choice of techniques should consider how the information contained in the observed data can be used, the accuracy of predictions, the amount of computational effort required, and how flexible they are in allowing both theoretical and empirical information to be utilized (Haining 1990).

Regression can be extended by incorporating a spatial autocorrelation parameter and by using different forms of the model with spatially lagged variables. A fair amount of trial and error is required to determine the best fitting model for a particular data set. For a particular study, the different regression models should be compared not only in terms of included and excluded variables but also in terms of different forms of the model. Interpretive knowledge can be utilized in the choice of model.

Regression assumes normality and stationarity, which may or may not hold. While regression is useful for estimating population means and totals, it does not always preserve the underlying relations among inventory attributes. For example, if one or more variables with missing values is difficult to predict, the correlation structure may be distorted. Regression may produce unreliable results for cases of high unique variances.

While traditional methods of interpolation (e.g., inverse distance weighting, local sample means) are less accurate, they are also less computationally intensive than kriging and do not require the subjective estimation of the variogram. Kriging assumes that the variogram is accurate over the entire study area. This is known as the stationarity hypothesis and in cases where it does not hold, the results from kriging will be less accurate. In addition, kriging assumes that the variable being predicted is multivariate normal. Raw data are rarely univariate normal (Rossi *et al.* 1994).

While MSN analysis is a robust approach that implicitly captures spatial

structure, it sensitive to sample size in that smaller samples sizes reduce the variance, which in turn affects it's predictive function. Results from MSN analysis are unpredictable if a strong non-linear relation exists among the variables of one or both sets. Canonical analysis (upon which MSN is based) is adequate for the case where the joint distribution of variables is linear and continuous. So, in cases where data are either heterogeneous or exhibit non-linear relations between variables, MSN will fall short. As well, the linear orthogonal nature of solutions in canonical analysis may be unrealistic for ecological data where there is often high unique variances.

An issue which is important but has not been addressed in this paper is that of the scale of variation. Spatial variation may be conceptualized as being composed of several (possibly independent) components at several different scales (Haining 1990). In testing for, representing and modeling spatial autocorrelation, it is important to identify and account for the contribution from each of the different scales to the total variation.

## CONCLUSIONS

If we rely on prediction models that assume that biological features are distributed uniformly or randomly in space, then chances of obtaining accurate predictions of missing values are small, since spatial autocorrelation characterizes most ecological data. It is often difficult to determine if assumptions have been broken, making it hard to compensate for such violations or to know how reliable the newly completed data set is. Consequently spatial autocorrelation must be tested for and techniques which either explicitly or implicitly consider spatial autocorrelation must be used when it is present.

While the different techniques examined above have various advantages, disadvantages and restrictions, it is impossible to produce a truly meaningful comparison unless each method is applied to a unique set of variables. To date there have been few studies which have formally compared the various spatial inference techniques which can be used to fill in missing data values. Given the increased importance such methods will have in the future, conducting such a comparative study is certainly warranted.

## REFERENCES

- Antle, A. N. and Klinkenberg, B. 1995. Statistically-based data generation techniques: an emerging trend. In: Heit, M., Parker, H. D. and Shortreid, A. (eds.) GIS Applications in Natural Resources 2, GIS World, Inc.
- Anselin, L. & Getis, A. 1992. Spatial statistical analysis and geographic information systems. *The Annals of Regional Science*, Springer-Verlag, Vol. 26, pp. 19 -33.
- Chrisman, N.R. 1991. The error component in spatial data. In: MacGuire D.J.,

- Goodchild, M.F. & Rhind, D.W. (eds.) Geographical Information Systems: Principles and Applications, Longman Scientific & Technical, Vol. 1, NY.
- Cressie, N. & ver Hoef, J.M. 1993. Spatial statistical analysis of environmental and ecological data. In: Goodchild, M.F., Parks, B.O. & Steyaert, L.T. (eds.), Environmental modeling with GIS, pp. 404 - 421. Oxford University Press Inc., Oxford.
- Goodchild, M. F. 1986. Spatial autocorrelation. CATMOG 47, Geo Books, Norwich.
- Goodchild, M. & Gopal, S. 1989. Accuracy of spatial databases. Taylor & Francis, Ltd., London.
- Griffith, D.A. 1987. Spatial autocorrelation : A primer. Monograph of the American Association of Geographers, Washington.
- Guptill, S.C. & Morrison, J.L. 1995. Looking Ahead. Morrison, J.L. (eds.) Elements of Spatial Data Quality, Elsevier Science, Oxford, UK.
- Haining, R. 1990. Spatial data analysis in the social and environmental sciences. Cambridge University Press, Cambridge.
- Legendre, P. & Fortin, M.J. 1989. Spatial pattern and ecological analysis. Vegetatio, Vol. 80, pp. 107 - 138.
- Legendre, P. 1993. Spatial autocorrelation: trouble or new paradigm? Ecology, Vol. 74, No. 6, pp. 1659 -1673.
- Moer, M. and Stage, A. R. 1995. Most similar neighbor: an improved sampling inference procedure for natural resource planning. Forest Science, Vol. 41, No. 2, pp. 337-359.
- Rossi, R.E., Mulla, D.J., Journel, A.G. & Franz, E.H. 1992. Geostatistical tools for modeling and interpreting ecological spatial dependence. Ecological Monographs, Vol. 62, No. 2, pp. 277 - 314.
- Rossi, R.E., Dungan, J.L. & Beck, L.R. 1994. Kriging the shadows: geostatistical interpolation for remote sensing, Remote Sensing Environment, Vol. 49, pp. 32 - 40.

### BIOGRAPHICAL SKETCH

Alissa N. Antle is a doctoral student in the Department of Geography at the University of British Columbia. She graduated from the University of Waterloo with a B.A.Sc. in Systems Design Engineering and a B.A. in Liberal Arts. Alissa's research interests include visualization and exploration of spatial data quality and using GIS for decision making in ecosystem management environments.

Peter L. Marshall is an Associate Professor in the Department of Forest Resources Management at the University of British Columbia. He received B.Sc.F. and M.Sc.F. degrees in forestry from the University of Toronto and a Ph.D. in forest management from the University of British Columbia. His research interests include sampling designs for collecting resource data and approaches to quantifying and predicting stand dynamics.

# Optimizing Sampling Schemes for Mapping and Dredging Polluted Sediment Layers

L. Hazelhoff<sup>1</sup> and F. Hoefsloot<sup>1</sup>

---

**Abstract.**-An optimal sample scheme of measurements of sediment depth was constructed for mapping and dredging a polluted sediment layer, taking into account economic constraints. The layer in question was deposited in a lake near Kampen in the Netherlands and consists of sediment discharged by the Rhine during the period 1932 -1980. A stochastic approach was chosen to estimate the depth of dredging that would result in a given amount of polluted sediment remaining in the lake after dredging has been carried out. The dredging result, i.e. the amount of sediment that is not removed, depends on the probability distribution of the estimated depth and the difference between the mean of this distribution and a chosen dredging depth. Two independent sources of error determine this probability distribution: i) the interpolation error and ii) the precision of the determination of the depth of dredging. The two sources of error are independent. Spatial interpolation was done using ordinary point kriging. The kriging error depends on the sampling scheme and the spatial variability of the depth of the bottom of the polluted sediment layer. Spatial correlation at three different scales was modelled with a nested variogram. The sampling distance of the sediment depth measurements was evaluated as a function of dredging result and dredging depth precision. This information can be used to derive a relation between sampling costs and dredging result. The amount of unnecessarily dredged clean sediment was estimated and incorporated in the procedure for designing an optimal sample scheme.

---

## INTRODUCTION

In the period 1932 - 1980 polluted Rhine sediments were deposited in lake Ketelmeer, located near Kampen in the Netherlands. This polluted sediment now has to be removed from an area of about 30 km<sup>2</sup>. A pilot study was carried out in two small areas to get an impression of the spatial variability of the depth of the under side of this layer. The sampling was proceeded in stages, with information from the first area being used for sampling the second. Within the pilot areas different sample stages were also applied.

<sup>1</sup>*Physical Geographers, University of Utrecht, The Netherlands.*

Dredging companies have recently used triangulation (TIN) to construct a digital model of the level of dredging, and in some cases the spatial variability was estimated using spectral waves or ARIMA models. The results of these studies were used to increase the depth of dredging below the TIN surface to ensure that downward fluctuations in the level of the polluted sediment will be removed. The sampling distance was arbitrarily determined. The main objective of this study is to estimate the extra dredging depth with reference to an interpolated DEM using Kriging and the Kriging variance. Using the Kriging variances, a relation can be established between extra dredging depth, sample distance and the amount of polluted sediment that is not removed by dredging. A second objective is to extrapolate the spatial variability of the polluted layer in two pilot areas to the whole lake so that economic assessments can be made of the sample scheme needed for the lake for a given fraction of polluted sediment that is not removed. A problem in this area is non-stationary spatial variation caused by human activity in the past and today. Digging of clay and sand has resulted in pits that have mostly been filled with polluted sediment, and along the shipping channels vessels cause resuspension and redeposition of the sediment along the channels. Both disturbances must be identified as outliers in the data, so careful exploratory analysis of the spatial data is necessary. First the spatial variability of the two sample areas is explored, then the method of determining the economically optimal dredging in relation with environmental constraints is explained and applied on the results of the second sample area.

The study was funded by the Ministry of Public Works (RWS).

### **Area one**

#### **sample scheme**

This area is 800x800 m and was chosen by RWS as a test location for redevelopment using new methods of dredging. Based on information from a survey on a slightly different sediment layer in the same area the following sample scheme was designed. Four subareas of 100x100 m were randomly chosen containing sample grids of 20 meter (figure 1). This approach can be justified by the knowledge of the sedimentation process; it is assumed that the spatial variability at this scale in the four selected blocks will be representative for the whole lake. To investigate whether the subareas are influenced by variation at a larger scale, a 100 meter grid was sampled in the whole area. A third level of scale was investigated at sample distances ranging from 0.5 to 5 m, realized by three groups of two short transects designed with the shape of a cross. The main objective of this short distance sampling was to test measurement accuracy; at these distances spatial variation was assumed lower than expected measurement error, RWS planned this lowest level of sampling. The information was also used to make a better estimate of the nugget.

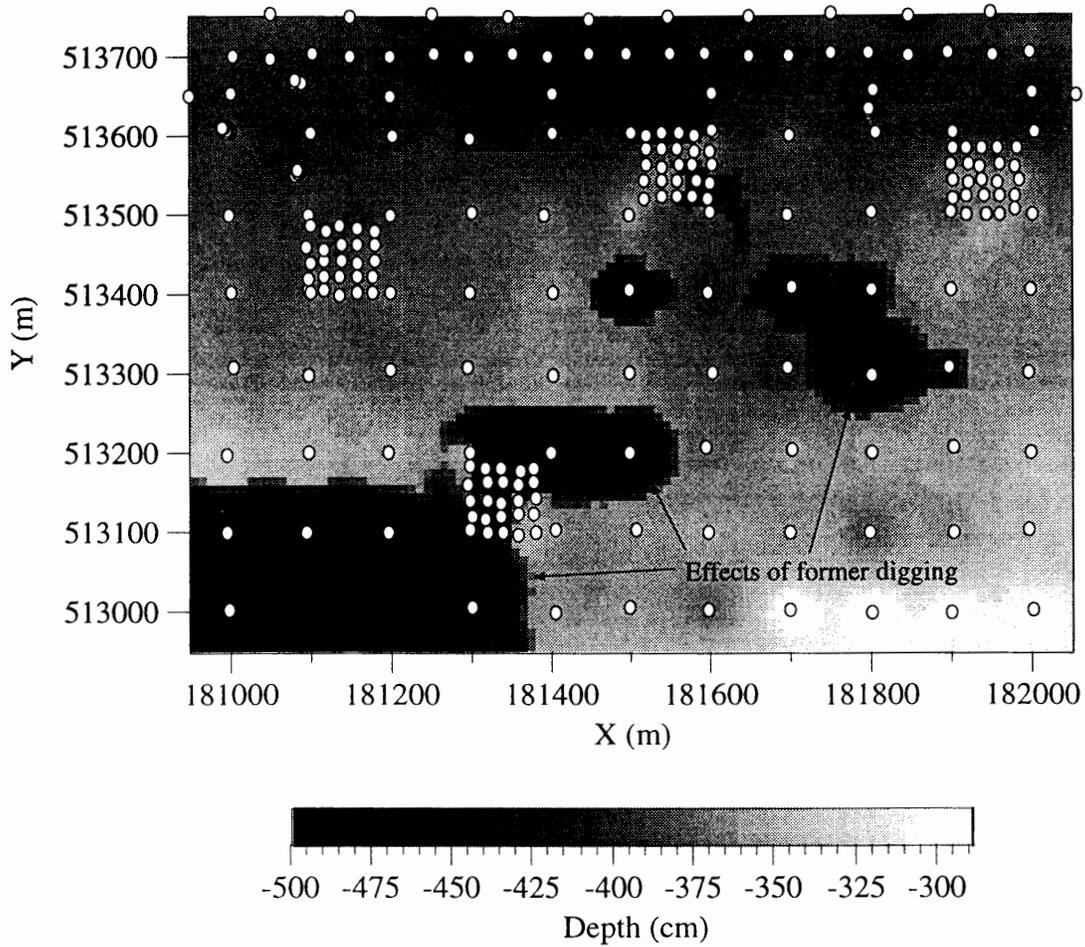
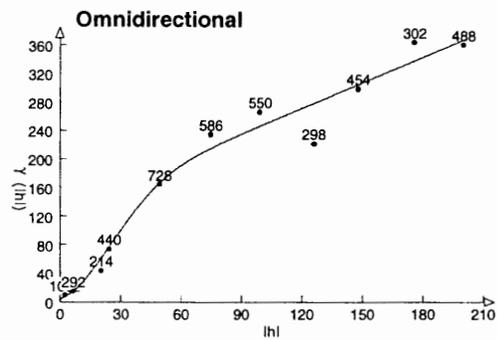


Figure 1. Sample design and interpolated surface of area one.

$$\text{Gamma (h)} : 4 + 124 \cdot \text{Gauss } 63 \text{ (h)} + 1.2 \cdot \text{Pow } 1 \text{ (h)}$$



figur 2. Semivariogram of the under side of the polluted layer in area one.

## **results**

In this area variation over distances of 100-500 m is clearly present with two of the subareas being located in zones with a trend. Performing an analysis on residuals reveals the variability at the 20-100 m scale, but spatial outliers were found in these subareas, using exploratory data analysis (EDA) software. The EDA software visualizes spatial variability by connecting measurement locations to lines which can be selected from a graph displaying the variogram cloud (Hazelhoff & Gunnink, 1992). These spatial outliers were identified as being caused by human interferences, and so they could not be used for investigating spatial variation at the 20-100 m scale. The variogram in figure 2 is based on 2 x 25 samples in this range. A nested variogram has been constructed; the experimental variograms for three spatial scales were calculated separately and then combined for fitting. It shows a low nugget caused by the limited spatial variation at the level of the crosses (0.5 to 5 m). A steep slope between 0 and 60m is clearly present.

## **Area two**

### **sample scheme**

In this area measuring 400x500 meter (figure 3) a dredging test was planned, so a DEM had to be constructed. A grid mesh of 50m was planned, based on the variogram of area one with the steep increase in the range of 0 - 60 m, so much lower kriging variances will occur when a 50 m grid mesh is used instead of 100 m. Due to practical constraints the realized sample scheme is somewhat different. In this area one subarea of 100x100 m was sampled at distances of 20 m located on a grid to compare the results with area one. Furthermore two crosses (with the same configuration as in area one) were added with sampling distances ranging from 0.5 to 5 m. A second sampling has been carried out for validation of the DEM. These samples also are plotted in figure 3, many of them are located close to the samples taken in the first stage.

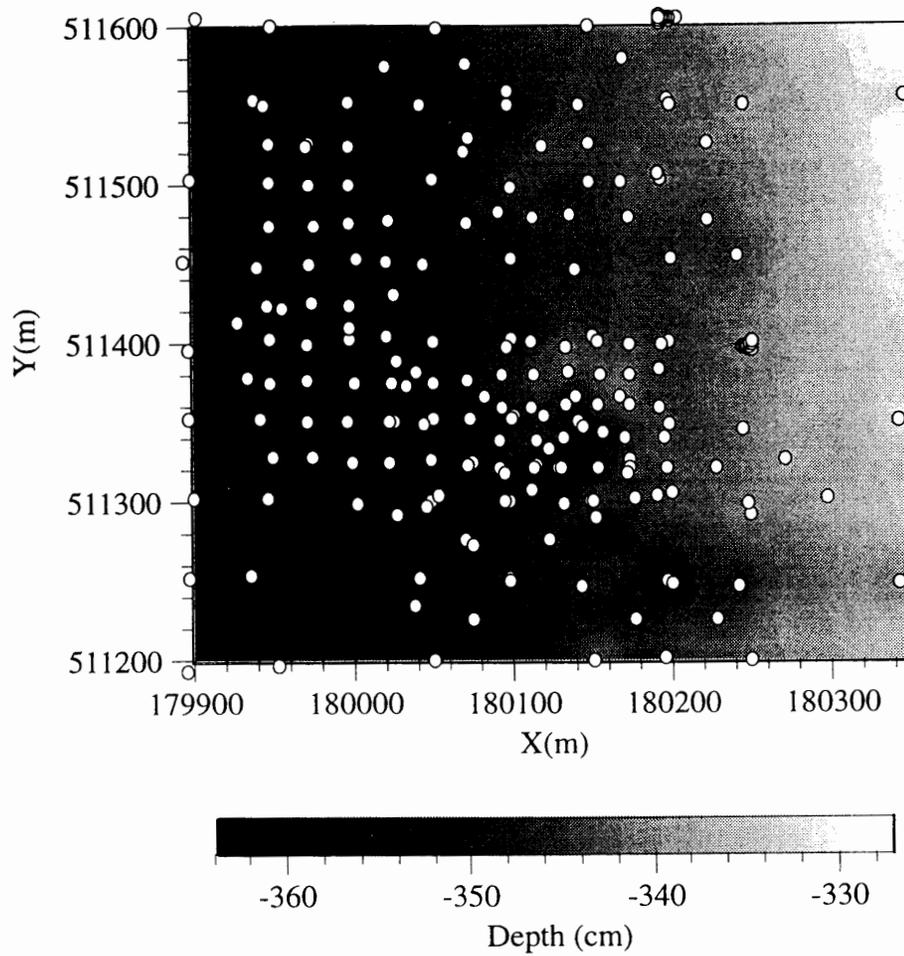


Figure 3. Sample scheme and interpolated surface of area two.

$$\text{Gamma (h)} : 9 + 10 \cdot \text{Sph } 30 \text{ (h)} + 0.013 \cdot \text{Pow } 1.53 \text{ (h)}$$

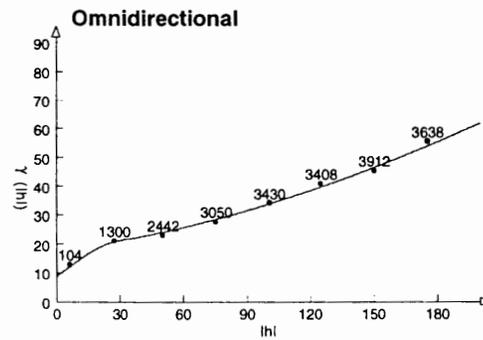


Figure 4. Semivariogram of the under side of the polluted layer in area two.

## results

After using EDA to identify human induced spatial outliers, the spatial variability appeared to be much lower than in areas one. The second stage of sampling for validation confirms the originally variogram, except the nugget variance appears now too low. Pairs of samples close together (<10 meter) and are not in the crosses show much higher semivariances than samples within the crosses. Two possible reasons can be given: 1) the crosses are located in areas with low spatial variability at short distances, and 2) measurement errors are lower within the crosses than outside the crosses. The variogram based on all samples is showed in figure 4. The steepest part ranges from 10 to 30 m, which is the important interval for kriging the 50 meter grid. The right part of the variogram fitted with a power model represents some linear trend in this area. For coarser grids this trend will be important.

### Dredging depth & residual sediment

The DEM of the lower side of the polluted layer in area two was made using ordinary point kriging (OPK). During the dredging tests the dredging depth was set at the level of the interpolated DEM which means that polluted sediment is still present (residual sediment) in half of the area, assuming a normal distribution of the deviations. The volume of the residual sediment can be calculated if the distribution of the deviations is known. In this study the kriging variance (normal distributed) and the error of estimating the dredging depth are used to describe the deviations from the DEM. These two sources of error are supposed to be independent, simply adding the variances gives the total error. Increasing the dredging depth relative to the DEM results in smaller volumes of residual sediment. The amount of residual sediment is calculated using the tail of the normal probability function. For example, if  $\sigma = 1$  cm then increasing the dredging depth by 1 cm results in a coverage of 16% of the area with polluted sediment. The residual volume  $V$  per unit of area can be calculated using (1).

$$V(z_d) = \int_{z_d} (z - z_d) p(z) dz \quad (1)$$

$z$  is the depth relative to the DEM where the center of the distribution lies and  $p(z)$  is the probability density of the normal distribution. Integration starts at dredging depth  $z_d$ . This integral is discretised to calculate  $V$ . The quantity  $V$  depends on  $\sigma$ , the kriging standard deviation (KSD) and  $z_d$ . KSD depends on the variogram and the sample scheme. The dredging result is defined as the result of dividing the residual volume by the total area.

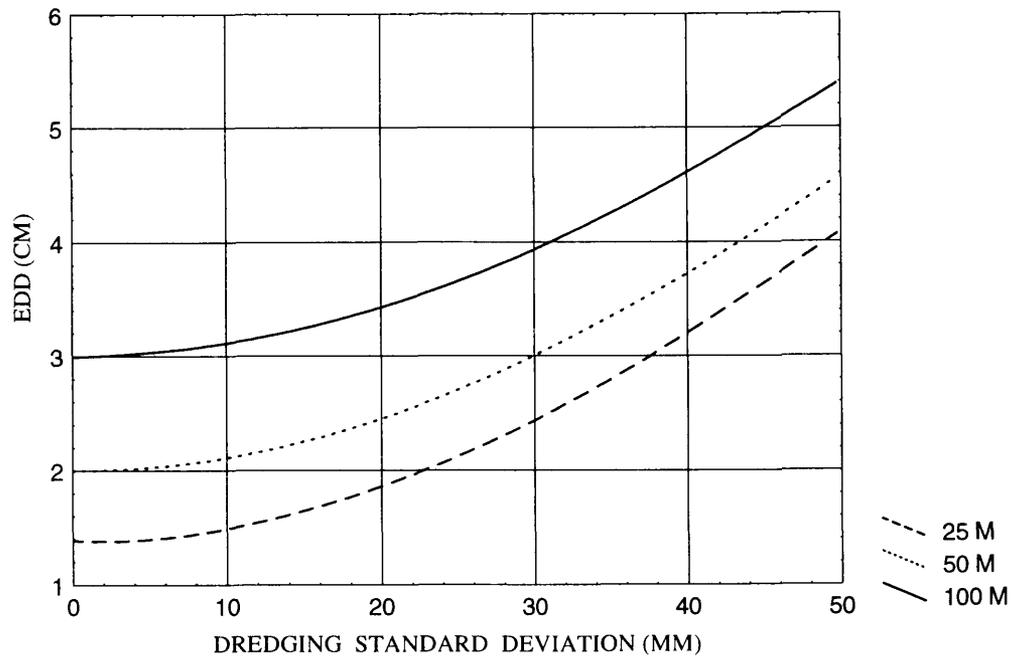
For area two the extra dredging depth (EDD) was calculated for a range of dredging results, and is repeated for different values of the dredging depth error. During the calculation each grid cell gets a different value for  $V$  because KSD varies over the area. A special computer program was written to work out this iterative procedure. Another interesting quantity is the volume of clean sediment dredged as function of EDD. Dredging at DEM depth means that dredged clean sediment (DCS) is equal to  $V$ . The volume of dredged polluted sediment is equal to  $V(z_d=z_{DEM}) - V(z_d)$ . The volume of clean sediment is calculated from formula 2. EDD is here expressed as a volume with an area of 1.

$$DCS = EDD - (V(z_d=z_{DEM}) - V(z_d)) + V(z_d=z_{DEM}) = EDD + V(z_d). \quad (2)$$

It is obvious from (2) that increasing EDD at low values of  $V(z_d)$  yields small improvement in the dredging result relative to the increase of DCS.

### **DREDGING RESULT & SAMPLE SCHEME**

For economic reasons the most interesting question is how does the EDD depend on the sample distance (regular grid) given a certain dredging result and a given the spatial variability described with a variogram? For the variogram of area two this relationship was calculated and plotted in figure 5. For three sample distances (100, 50 and 25 m) the relation between dredging standard deviation DSTD and EDD is plotted. The dredging result is set at 1 cm, which is a practical value. DSTD is an important error source because the values are comparable with the KSTD based on the three sample distances. Improvement of estimating dredging depth results in a significantly lower EDD. At the moment some dredging companies claim values for DSTD near 2 cm. It is up to RWS to balance increasing cost due to extra dredging depth against the increasing cost of denser sampling.



**Figure 5. Relation between dredging standard deviation and extra dredging depth with a given dredging result of 1 cm.**

### **Extrapolation from pilot areas to the whole lake**

The relations between sample density and EDD were derived from the variogram of area two. The variogram in area one differs considerable from this variogram; in area one kriging variances are larger for any given sample distance. But the spatial variability in area two is most likely to be the most representative. Area one is located near the border of the lake with a lot of human disturbance. If it is true that area two is representative for the lake, a 50 m grid uses information from the spherical part of the variogram. The recommendation to RWS is to start with a 100 m grid to explore the spatial variability at a larger scale than 50 m, as is clearly present in the two areas. In areas with such a large spatial variability that a large EDD is needed to match the desired dredging result, the 50 m grid is advisable. Another important fact is the presence of human disturbances which need dense sampling to be tracked. There is a risk that disturbances of at most 1 ha will be missed using a 100 m grid, which can be important because the thickness of the polluted sediment in the areas may be more than doubled.

## **CONCLUSIONS**

Spatial correlation at two distinctive scales separated at about 40 meter is present in the lower boundary of the polluted sediment.

Ordinary point kriging can be used together with an error specified for estimating dredging depth, to optimize sample schemes given a dredging result. This information cannot be derived from TIN models and associated approximators of variability like ARIMA and spectral waves.

For the spatial variability in the second pilot area, dredge depth accuracy is an important source of error.

## **REFERENCES**

Hazelhoff, L. & Gunnink, J.L. 1992. Linking Tools for Exploratory Analysis for Spatial Data. In Harts, J., Ottens, F.L.; Conference Proceedings EGIS 1992.

## **BIOGRAPHICAL SKETCH**

Lodewijk Hazelhoff is a Physical Geographer specialized in environmental assessment and GIS. He graduated from the University of Amsterdam in 1980.

Frans Hoefsloot is a Physical Geographer specialized in environmental assessment. He graduated from the University of Utrecht in 1992.

# Forest Cover Monitoring in India : The Satellite Experience

**J.B. Lal<sup>1</sup>**

---

**Abstract :** In certain quantitative aspects, India is the biggest user of satellite imagery in the monitoring of a natural resource. The Forest Survey of India produces a biennial forest cover map of the entire country based on interpretation of satellite imagery. The interpretation which was in the beginning manual, is now a combined digital and visual effort. The area covered in biennial mapping is 328 million ha. The average annual coverage works out at 164 million ha. Apart from the resolution of the sensor, the scale of mapping significantly influences the accuracy of maps. The Forest Survey of India has so far produced four series of maps. The first series was produced in 1987, and the last in 1993. The 1987 maps were on 1: 1 million scale based on visual interpretation of the Landsat MSS imagery. The final series is on 1:250,000 scale based on interactive interpretation of Landsat TM and IRS LISS II imagery. From the first to the fourth of the series of the maps there is a big improvement not only in accuracy but also in the utility of the maps.

The first series of the maps was useful only in strategic planning at national and state level, while the fourth series is useful also in operational planning at district (the fundamental administrative unit of India) level.

---

## INTRODUCTION

A few years from entering the twenty-first century, India still lives in a forest of biomass era. It still uses over 200 million cubic meter of fire-wood for domestic energy and still lets over 90 million cattle graze in forests. One-sixth of its population, recorded as scheduled tribes in its constitution, still meets most of its living needs from forests. "(LAL 1989a)". Forests are vital to India not only for ecological, but also for socio-economic reasons. Consequently, rapid, regular and reliable forest cover assessment and monitoring is an indispensable activity for India.

---

<sup>1</sup> *Former Director, Forest Survey of India*

Unfortunately, till nineteen-eighties, when satellite data was first made use of in forest cover assessment, India, at no point of time, had upto date information about the status of its entire forests. All it knew was the extent of area legally classified as forests.

Synoptic, repetitive, and digital as it is, satellite data could not have an alternative to itself in forest cover monitoring in a large and diverse country like India, where biotic pressures change the situation of a natural resource in months, let alone years. Even on 1: 50,000 scale the country is covered by 93,000 aerial photographs, while only in 200 sheets of satellite imagery from Landsat on 1:250,000 scale the entire country could be viewed. "(Lal 1989b)". Further, in India which contains as many as 221 ecologically stable forest types which vary from tropical wet evergreen forests deep in the south to alpine vegetation in the north in the Himalayas, and from dry thorn forests in the extreme west to moist deciduous forests on the eastern border "(Lal 1989b)", multispectral and multi-temporal data is of paramount importance.

Apart from technical reasons, satellite data also saves time and money in obtaining adequate results in regard to forest cover assessment and monitoring.

### **FOREST COVER MONITORING IN INDIA**

Using the satellite data the Forest Survey of India (FSI) has so far made four biennial efforts in assessing forest cover, and preparing forest vegetation maps of the country. In its vegetation mapping, FSI "(1988,1989,1993,1994)" recognised only four categories of the land cover :

dense forest : forest of 40% crown density or more,  
open forest : forests of less than 40% but more than 10% crown density  
scrub forest : Tree-lands with less than 10% crown density  
mangrove forest : -  
non-forest lands : -

The scale of mapping, the source of satellite data, the sensor, and other related information in regard to four assessments is summarised in Table 1:

**Table 1**

Assessment year	Name of the Satellite	Name of the sensor	The period to which imagery related	Resolution	Scale of mapping
1987	Landsat	Multi-Spectral Scanner(MSS)	1981-83	79m	1:1000,000
1989	-do-	Thematic Mapper(TM)	1985-87	30m	1:250,000
1991	-do-	-do-	1987-89	-do-	-do-
1993	Landsat and IRS	TM and LISS II	1989-91	36m	-do-

Source : FSI (1988,1989,1993,1994)

The interpretation of the imagery in regard to first three assessments was purely manual, but the last was a combined visual, digital effort. The results obtained in the four assessments are summarised in Table 2 :

**Table 2**  
(in million hectare)

Assessment year	Forest			Scrub	Non-forest	uninte- rpreted (clouds etc.)	Total land area
	Dense	Open	Man- grove	Total			
1987	36.14	27.66	0.40	64.20	7.68	255.74	1.15
1989	37.85	25.74	0.42	64.01	6.61	257.77	0.39
1991	38.50	24.99	0.42	63.91	5.97	256.93	1.92
1993	38.56	25.03	0.42	64.01	5.90	258.82	Nil

Source : FSI (1988,1989,1993,1994)

It may, however, be mentioned that the comparison of 1987 assessment with other three assessments is not totally valid. First, in the 1987 assessment the forest cover maps were prepared on 1:1million scale while in other assessments on 1:250000 scale. There is a big difference in the cartographic limits of the two scales. The minimum area that can be shown on 1:1 million scale is 400 ha while that in 1:250,000 scale it is 25 ha. In other words, in the first assessment, the isolated patches of forests less than 400 ha in extent were omitted from the counting. In the same manner the enclaves of other land-uses measuring less than

400 ha in forested region were counted as forests. Secondly, the imagery used for 1987 assessment was from Multi Spectral Scanner(MSS) which has a spatial resolution of 79m while in other assessments the imagery of Thematic Mapper(TM) which has a spatial resolution of 30m was used.

### **ACCURACY OF FOREST ASSESSMENTS**

A fact which needs to be categorically mentioned is that in classifying forests in 'closed' and 'open' does not reveal the true status of forests in regard to composition and crown density. The crown density of a forest may reduce from 80 to 50%, the forest would still be classed as 'dense'. And, due to human intervention a forest may become less diverse in species composition, but the satellite imagery would say little about it.

It should also be appreciated that the difference in radiance value of a land unit at two points of time may be due to several factors, such as :

- \* change in vegetation
- \* differences in illumination/sun angle
- \* differences in atmospheric conditions
- \* differences in ground moisture
- \* differences in sensor calibration

Therefore, radiance change on account of change in crown density should be relatively large for imagery to give authentic information about land cover.

Though forest vegetation maps prepared by FSI in various assessments were subjected to extensive ground truth verification, no systematic accuracy assessment exercise was undertaken by the organisation. Lal and Prajapati(1989), however, carried out a reliable exercise in Nagaland state to ascertain the effect of scale on the accuracy of vegetation mapping from satellite data. They reported that over-all classification accuracy in case of vegetation mapping on 1:1 million scale from MSS data was only 54%, while on 1:250,000 scale from TM data it was 79%.

The FSI in an informal communication to the author have reported that digital interpretation of satellite data has increased the classification accuracy to 90%. The location error of a point did not go beyond 30 m.

### **GRASSLAND ASSESSMENT**

Apart from assessing and monitoring forest vegetation, the space-borne data has not been made much use of in assessing and monitoring other land uses. A paramount need in the country which has over 400 million cattle, is the

assessment of the extent and status of grasslands. The Space Application Center of India has recently undertaken the work of mapping the grasslands of the country by use of the satellite data. The work is yet to be completed. However, Lal, Gulati and Bist (1991) did a limited exercise of mapping the alpine pastures in the Himalayas. The results of the exercise are summarised in Table 3.

**Table 3**

Name of the State	Geographical area ( in million ha)	Area of alpine pasture
Arunachal Pradesh	8.36	1.23
Sikkim	0.73	0.16
Uttar Pradesh	29.44	0.85
Himachal Pradesh	5.57	1.73
Jammu and Kashmir	22.22	13.19
Total	66.32	17.16

Lal et al (1991) assessed the interpretation accuracy by ground-truth verification of sample points. The number of sample points was worked out by using the formula

$$n = \frac{p \times q}{d^2}$$

where n=number of sample points

p=desired proportion of estimated accuracy

q=(100-p)

d=standard error acceptable at desired confidence level

p,q,d are all expressed in per cent.

The sample points were selected by random sampling out of a population formed by 12 x 12 cm grids on the map of alpine pastures. Though it was beyond expectation, the interpretation accuracy was found to be 100%.

## UTILITY OF FOREST VEGETATION MAPS

Apart from being of great value in strategic planning at the national level, forest vegetation maps prepared from the satellite data have been used in tactical planning at the state level. From the third assessment "(FSI,1993)" onward the forest vegetation maps have been prepared district-wise. A district is the fundamental administrative unit in India, and forest vegetation maps' being available every two years for a district would facilitate operational planning.

Forest vegetation maps obtained from satellite data gives information not only on change in land cover, but also on change in land use. In 1989, Forest Survey of India "(Lal,1989b)" carried out an exercise to assess the change in area affected by shifting cultivation in north-eastern hill states of India. The exercise indicated that during the period 1974 to 1984, area affected by shifting cultivation decreased from 73410 ha to 62854 ha. The State of Forest Report,1993 "(FSI,1994)" makes a useful analysis of the losses and gains in forest cover in the north-east hill states of India due to change in land use. The report points out that there was a loss of 70200 ha of forest cover on account of more forest area being brought under shifting cultivation, while there was a gain of 31500 ha in forest cover on account of natural regeneration of abandoned shifting cultivation sites.

## CONCLUSIONS

The utility of satellite data in land-cover and land-use assessment and monitoring has been proved beyond doubt. The vegetation maps prepared from satellite data are fairly accurate as the emperical testing would indicate, but a fool-proof statistical design for accuracy assessment has yet to be evolved for Indian conditions which are ecologically so diverse. For interpretation accuracy, two points have to be specially kept in mind First, in regard to decidous forest it has to be ensured that imagery pertaining to the period trees are leafless to any degree is not selected for interpretation . Secondly, elaborate experiments have to be organised to test whether, as Lal has pointed out "(Lal,1989b)", reflectance recorded by satellite sensors, is more a function of 'leaf area duration' than of 'leaf area index'.

## REFERENCES

- Forest Survey of India (FSI). 1988. State of Forest Report,1987. 7-20pp.
- Forest Survey of India. 1989. State of Forest Report,1989. 7-11pp.
- Forest Survey of India. 1993. State of Forest Report,1991. 7-11pp.
- Forest Survey of India. 1994. State of Forest Report,1993. 7-11pp.
- Lal,J.B. 1989a. India's Forests : Myth and Reality, 2nd ed. Dehradun :  
Natraj publisher 44,89-97 pp

- Lal, J.B. 1989b. Forest Resource Management from Space. Van Vigyan. Vol. 27, No. 4. 1989. 168-178pp.
- Lal, J.B., and R.C. Prajapati. 1989. Effect of Scale on Vegetation Mapping by Remote Sensing. The Indian Forester. Vol. 115, No. 6, June, 1989. 372-377pp.
- Lal, J.B., A.K. Gulati and M.S. Bist. 1991. Satellite mapping of alpine pastures in the Himalayas. Int. J. Remote Sensing. 1991, Vol. 12, No. 3. 435-443pp.

### **BIOGRAPHICAL SKETCH**

J.B. Lal, a former member of Indian Forest Service, is an Associate of Indian Forest College, and a Ph.D from the Victoria University, Manchester, U.K. He has held top forestry positions in the Govt. of India. He was Director, Forest Survey of India, Director General, Indian Council of Forestry Research and Education, and Director, Indian Institute of Forest Management. He has served as a member of National Environmental Council of India, and as a consultant to FAO and SIDA. He has written three books, and several technical papers.

# Study About the Use of SPOT Data for the Estimation of Pastoral Areas in Morocco

Philippe Brion<sup>1</sup>

---

**Abstract.**-In order to provide information on land use for a whole country, it is often impossible to have a complete cover of the land with remote sensing data, and it is necessary to use a sampling method. This paper presents an approach of this problem, relatively to the estimation of pastoral areas in Morocco : selection of SPOT images, and then selection of "segments" inside the images. The different options of the sampling plan are discussed.

---

## INTRODUCTION

In order to provide information on land use, remote sensing brought many hopes. In some cases, it has been used jointly with ground surveys, and the exhaustive cover of the satellite image has given an auxiliary information, combined with the results of the ground survey in order to produce an improved estimate (by the method of the regression estimate, see for example Pastorelli, 1992). But, often, the number of images required for having an exhaustive cover is too high, and financial reasons lead to the use of a sample of images : this is, for example, the case of the MARS program (of the European commission, for the estimation of crop areas as cereals) ; however, there are no established rules for the sampling plan to be used for that kind of problems.

The object of this paper is to propose an a priori approach to this problem, concerning estimation of pastoral areas in Morocco. This work was done jointly with CNES (Centre National d'Etudes Spatiales, Toulouse) and CRTS (Centre Royal de Télédétection Spatiale, Rabat).

---

<sup>1</sup>Statistician, INSEE, Paris, France

## **PRINCIPLES OF THE PROPOSED APPROACH**

### **Constraints and method**

The first constraint concerns the purchase of satellite data (here, SPOT images). The impossibility to have a complete cover of Morocco leads us to have a sample of images. Then, since the classification of images does not presently give quite perfect results (there still remains confusions for some types of land use), we are led to having more complete investigations for portions of the images (by photointerpretation for example).

So, we propose a two-stage sampling plan for the estimation of pastoral areas:

- first, selection of SPOT images ;
- second, selection of "segments" inside the selected images.

These segments are "surveyed" by photointerpretation, or by a ground survey. They have, for practical reasons, a square shape.

Which items of the sampling plan are to be studied ?

- The size of the segments has to be chosen. In France, for example, a size of 50 hectares is used for ground surveys because of practical reasons (work of the enumerator). What should be proposed in the context of pastoral areas in Morocco?
- Another question concerns the number of segments to be selected inside a SPOT image.
- Third, the sample of segments may be selected using a systematic sampling method. Is this method efficient, and if so, what is the gain of precision to be expected ?
- The selection of SPOT images is another problem : how many should be selected, and is it possible to use a stratification of the country for the sampling plan ?

Then, once the former points have been treated, what global precision is to be expected ?

### **An experimental approach**

This approach consists in using some existing materials in order to produce, a priori, some quantified results to the former questions.

First, three SPOT images, belonging to regions of Morocco with different characteristics (in the upper Atlas, near the Sahara and on arid plateaux) have been classified, especially isolating the item "pastoral area" ; these images have been used for the study of the second stage of the sampling plan.

The study of the first stage was realized using data proceeding from a SPOT "Quick-Look-Numerical" (QLN) : these data are obtained by sampling one pixel for six, on row and column, and give a global view of the country, much bigger

than a single image. These data were also classified, particularly using the item "pastoral area".

The use of these materials permits to have quantified data to study the sampling system before the actual survey : this approach is rather a new one for the problem studied, because until now, we have had at our disposal experiences where the precision is evaluated a posteriori, once the survey has been conducted.

At last, it is also important to mention that the subject discussed in this paper is a statistical one, and not a cartographic one.

## RESULTS

First, we will discuss the results concerning the second stage (selection of segments inside the images), and integrate them to those obtained for the first stage (selection of SPOT images).

### Sample of segments inside the images

The three images classified were cut up in segments, using six different grids to define the segments : 200\*200m, 400\*400m, 600\*600m, 800\*800m, 1000\*1000m, 1200\*1200m.

#### Size of the segments

The variance of the estimate of the mean of a variable  $Y$ , using a simple random sample of  $m$  segments is approximately equal to :

$$V(\text{estimate of the mean}) = V(Y) / m \quad (1)$$

if  $V(Y)$  is the variance of the variable in the population, and if we do not take into account the sampling fraction  $f$  that is very small, so the quantity  $(1-f)$  is approximately equal to 1.

Table 1 displays the value of  $V(Y)$  for the variable "proportion of pastoral areas inside the segments", regarding the three images and the different sizes of segments.

Segment size	Midelt	Oujda	Ktaoua
200 * 200	0.132	0.078	0.165
400 * 400	0.111	0.067	0.144
600 * 600	0.102	0.059	0.128
800 * 800	0.095	0.057	0.117
1000 * 1000	0.091	0.054	0.108
1200 * 1200	0.087	0.050	0.100

When the size of the segment increases, the variance  $V(Y)$  decreases, because the bigger segments are more homogeneous. But, if we introduce cost elements

(corresponding to the time needed to "survey" the segments, e.g. proportional to the area of the segments, or at least to the square root of this area), we see that the decrease of variance is not sufficient enough to compensate cost increase. So, according to the results obtained for these three images located in different regions of Morocco, it seems better to use small segments (200\*200m).

### **Efficiency of the systematic sampling method**

This method consists in selecting segments located on a grid, rather than taking them completely randomly. The size of the grid may vary, and is related to the sampling fraction. Having cut up three images in segments makes it possible to calculate exactly the variance of the estimate with this method (by calculating the results of all possible systematic samples, which are in limited number), and to compare it to the one obtained by Eq. 1, for the three sites, different sizes of segments and for different sampling rates.

The results are that, generally, the systematic method is more efficient than simple random sampling, and, on an average, the gain of variance of the estimate is between 30% and 40%. These results are similar to those obtained by Dunn and Harrison (1993).

### **Precision of the estimation inside the images**

Table 2 gives the coefficient of variation for the estimate of the proportion of pastoral areas inside each image for a systematic sampling of segments of 200\*200m, the sampling fraction being 1/100.

**Table 2.-Precision of the estimates inside the images**

<b>Image</b>	<b>coefficient of variation</b>
Midelt	2.3 %
Oujda	1.1 %
Ktaoua	9.9 %

The coefficient of variation is defined as the ratio from the square root of the variance of the estimate of a characteristic to the value of this characteristic.

### **Selection of SPOT images**

Using the QLN makes it possible to reconstruct "approximate" SPOT images for the whole country. The values obtained from these images constitute the basis to evaluate the global precision of the sampling plan.

### **Variance of the estimate for a two-stage sampling plan**

If the selection for the first stage is with equal probabilities, the variance of the estimate  $\hat{Y}$  of the total is :

$$V(\hat{Y}) = \frac{M^2}{m} \left(1 - \frac{m}{M}\right) S_1^2 + \frac{M}{m} \sum_{\alpha=1}^M Z_{\alpha} \quad (2)$$

Where

$$S_1^2 = \frac{1}{M-1} \sum_{\alpha=1}^M (T_{\alpha}(Y) - \bar{T})^2$$

$$\bar{T} = \frac{1}{M} \sum_{\alpha=1}^M T_{\alpha}(Y)$$

$T_{\alpha}(Y)$  pastoral area of the image  $\alpha$

$Z_{\alpha}$  is the variance of the estimate of the total  $T_{\alpha}(Y)$  according to the method of selection of segments inside the image

Results from the first part (selection of segments inside the images) are used for the estimation of the  $Z_{\alpha}$ , and the data of QLN give information about  $S_1^2$ .

## Results

Average results (calculated with the three images) for the second stage were used, and Table 3 gives the precision of the global estimate of pastoral areas for Morocco, with different numbers of images selected at the first stage, and 400 segments of 200\*200m selected in each image.

**Table 3.-Precision of the estimate of pastoral area for Morocco**

Number of surveyed images	Coefficient of variation of the estimate (%)	Relative share of the second stage in the total variance of the estimate (%)
20	12.0	2.1
30	9.3	2.4
40	7.6	2.7
50	6.3	3.1
70	4.5	4.3
100	2.2	12.3

The share of the variance which depends from the second stage is small, that means that the variance between images brings the most important part of this variance. So, it is possible to have, for the second stage, a smaller sample. If we take 30 images and 30 segments inside each image, we obtain a coefficient of variation of 10.5%.

The second result of Table 3 is that, even when the sampling rate for the first stage is high, for example 100 images selected within the 115 "possible" which cover the whole country, there still remains a consistent value of the variance of the estimate.

### **Use of a stratification**

A "rough" stratification was used, classifying all 115 images in 9 classes : selecting 30 images using this stratification (and 30 segments inside each image), the coefficient of variation of the estimate for pastoral areas is 8.9% (instead of 10.5% without stratification).

## **CONCLUSION**

The most important result of this study is the great part of the variance of the estimate due to the first stage of the sampling plan, that is the fact of selecting images among the 115 images covering the country.

This result is well known for sampling questions, and one way to take it into account and to improve the efficiency of the method might be to have smaller primary units : using portions of SPOT images for the first stage could be a response to that problem.

Then, the results obtained in this paper should not be considered as exact values of the precision to be obtained with such a method, but rather as approximations based on an experimental approach.

Finally, we considered this problem as an "instantaneous" estimation, trying to evaluate the pastoral areas of the country for one year. To evaluate the evolution between two years, or more, should lead to other propositions and results for the sampling plan.

## **REFERENCES**

- Cochran, W.G., 1977, Sampling techniques, 3rd ed, Wiley, New York  
Dunn, R., Harrison, A.R., 1993, Two-dimensional systematic sampling of land use, Applied Statistics 42 n°4  
Pastorelli, R, 1992, Superficies agricoles à partir d'images satellite, Courrier des statistiques n°61-62, Paris  
Rapport CNES, 1995, Projet Géostat Maroc, étude statistique, CNES, Toulouse

## **BIOGRAPHICAL SKETCH**

Philippe Brion graduated from Ecole Polytechnique (Paris) and ENSAE (Ecole Nationale de la Statistique et de l'Administration Economique, Paris). He was in charge of the sampling methodology in SCEES (statistical department of the French Ministry of Agriculture) from 1980 to 1989, and is now working in INSEE (Institut National de la Statistique et des Etudes Economiques) on methodological problems for statistical cooperation (developing and transition countries).

# The Influence of Vegetation Cover Density and Topographic Parameters on the Thermal Emission of the Beech Forests of Simbruini Mountains (Central Italy)

Carlo Ricotta, Giancarlo Avena and Fernando Ferri<sup>1</sup>

---

**Abstract.** - Numerous studies have noted a strong negative correlation between radiometric surface temperature and spectral ratio-based vegetation indices. Since vegetation, through the process of transpiration, considerably enhances evaporative fluxes to the atmosphere, greater vegetation biomass should be associated with increased reduction in surface temperature. In irregular terrains with variable sun-sensor-surface geometry one of the major problems concerning the evaluation of surface temperatures of vegetation from remotely sensed data is that the thermal response of the earth surface is strongly influenced by topographic factors, such as altitude and insolation angle. In this paper, which uses remotely sensed inputs of surface temperature and vegetation fractional cover, the relations between the surface temperatures of the beech forests of Simbruini Mountains (Central Italy) and the main vegetation cover density and topographic parameters are discussed. According to the results obtained, 79.0% of the thermal response variance of the beech forests is due to variations of the topographic parameters.

---

## INTRODUCTION

Thermal infrared measures of the Earth's surface from terrestrial and airborne sensors have been used since the 1960s to investigate energy budget conditions of vegetation. With the launch of Landsat 4 and 5 plant ecologists now have access to a wide-area synoptic and multitemporal thermal infrared band TM6 with a

---

<sup>1</sup>Remote Sensing Laboratory, Department of Plant Biology, University of Rome "La Sapienza", Rome, Italy.

relatively high resolution (Lathrop e Lillesand 1987). A number of authors (Asrar et al. 1988, Smith and Choudhury 1990, Smith and Choudhury 1991, Hope and McDowell 1992) had focused on the widely observed negative correlation between radiometric surface temperature and remotely sensed measures of actively transpiring vegetation such as vegetation indices.

One hypothesis to explain this observation is that variability in transpiring leaf area produces spatial variability in surface temperature by modulating evaporative cooling at the land-surface-atmosphere interface (Friedl and Davis 1994). In mountainous regions, however, this relationship may be complicated by the effect of topography since a variation in altitude and insolation angle should also result in variated surface heating. The combined effect of topography and evapotranspiration in the energy exchange process at the surface produces a complex problem in the interpretation of observed thermal emissions (Conese et al. 1989). The present work aims at estimating by Landsat 5 TM data the variability of the beech forest surface temperatures located on the Simbruini Mountains (Central Italy) depending on the major remotely sensed vegetation cover density and topographic parameters, such as the Normalized Difference Vegetation Index (NDVI), altitude and insolation angle.

## STUDY AREA

The beech forests of the Municipalities of Trevi nel Lazio and Filettino (Central Italy) were selected for analysis. There are dense 5600 ha forests located in the watershed of the Aniene River with elevational extent from about 900 to 1700 m. Due to environmental importance, the area was designated in 1983 as part of the Natural Park of Simbruini Mountains.

The study area belongs to the Mountainous/Sub-alpine bio-climatic belt (Blasi 1994). The climate is strongly influenced by the position of the Simbruini Massive, which forms a transversal barrier to the western sea-winds. Meteorological data came from the station of the Ministero dei Lavori Pubblici located in Filettino (1062 metres above sea level.). The average annual temperature is 10.7 °C. The average annual rainfall is 1413.9 mm with a principal maximum in winter and a secondary maximum during fall. The average rainfall in the quarter June-August is 155.7 mm with a consequent lack of summer aridity. A short period of summer sub aridity is still possible during particularly dry years. The range of relief is approximately 1,600 metres.

The maximum elevation above sea level is 2,156 metres (MSL). The topography of the area is complex with a variety of site exposures. The beech forests represent a nearly continuous formation with high cover density values due to the absence of intense grazing and cutting activities. The dominant tree species is *Fagus sylvatica*. More rare arboreal species are *Acer pseudoplatanus*, *Taxus baccata*, *Tilia platiphyllos* and *Ulmus glabra*.

## MATERIALS AND METHODS

The study area was extracted from a Landsat TM 190/31 II Q scene from 29 July 1993. Collateral information consisted of the vegetation map of the study area at 1:25,000 scale obtained through photointerpretation of infrared aerial photographs. Minimal and maximal air temperatures were respectively 12.8 °C and 26.8 °C at Filettino on 29 July 1993. In the month prior to satellite overpass there was 79.4 mm of rain. 17 July was the last rainy day prior to satellite overpass with 70.0 mm of rain. On 29 July 1993 the atmosphere was clear and any atmospheric attenuation was assessed to be systematic across the study area.

The TM data were rectified to a Universal Transverse Mercator grid using a nearest neighbour resampling method. A digital terrain model (DTM) was created by acquiring on a digitizer the contour lines every 25 m of the national cartography mapping sheets at 1:25,000 scale of the study area. Relative sun incidence was calculated from the DTM. Relative sun incidence displays direct illumination which is equal to the cosine of the angle between the surface normal and the incident beam for each pixel. The solar azimuth and altitude angles at the time of satellite overpass (10:09 local time) necessary to calculate this illumination model were read from the header file of the computer compatible tape (CCT).

In order to test the hypothesis that thermal emissions are related to the topographic parameters, such as altitude and insolation angle and the amount of vegetative cover, a systematic sampling (Congalton 1988) of 216 pixels was performed on the vegetated areas of the image. To avoid autocorrelation effects particularly on band TM 6 the starting point was located at random and each successive unit was taken at an interval of 10 lines and 10 columns thereafter.

For each of those pixels the values of altitude (MSL), relative sun incidence ( $\cos Z$ ), NDVI ( $(TM4-TM3)/(TM4+TM3)$ ) and surface temperatures were extracted from the data set.

## THERMAL DATA CALIBRATION

The satellite-derived normalized difference vegetation index (NDVI) is used as the index for the vegetation density. The physical quantity that is measured remotely is the spectral radiance reaching the sensor [ $mW/cm^2 sr \mu m$ ] quantized in digital numbers (DN) over a 8-bit brightness range. Digital counts of the bands TM3, TM4 and TM6 were converted to spectral radiances according to the method proposed by Markham and Barker (1986).

The improved darkest object subtraction technique described by Chavez (1988) was used to correct digital numbers from the red and near-infrared bands for differences in atmospheric scattering in these two bands. This simple correction was applied to minimize compression in the NDVI values which results from

unequal atmospheric effects on the red and near-infrared radiances (Hope and McDowell 1992). Spectral radiance  $L$  in the thermal-infrared band and surface temperature  $T_s$  are related in the form (Singh 1988):

$$\ln(L) = a + b/T_s \quad [1]$$

Note that the accuracy of Eq. 1 decreases with the increase of the range of temperature chosen. For this reason the parameters  $a$  and  $b$  were determined for temperature range of 40 °K only and are reported by Singh (1988).

This technique does not account for atmospheric attenuation of the thermal radiance. However, although the 8-14  $\mu\text{m}$  atmospheric window is far from being transparent, Landsat TM thermal IR data collected under clear atmospheric conditions can be correctly calibrated in surface temperatures with acceptable accuracy even when corrections for atmospheric effect are not applied. This is possible because under clear atmospheric conditions the amount of atmospheric path radiance is nearly equal to the amount of attenuated target radiance (Bartolucci et al. 1988). Moreover, since the objective of this study was to investigate the correlation coefficients of the relationships between  $T_s$  and the examined independent variables, a systematic offset in the values would not affect the interpretation of the results. The relationship between the surface temperature of the beech forests ( $T_s$ ) and the three independent variables MSL, COS Z and NDVI were investigated. Linear correlation coefficients were calculated to quantify the magnitude of the hypothesized relationships, and scattergrams were constructed to indicate the nature of the relationships.

## RESULTS AND DISCUSSION

Scattergrams depicting the relationship between surface temperatures of the beech forests and the independent variables are presented in figure 1 and 2, and the associated correlation coefficients are given in table 1.

**Table 1. - Correlation coefficients (R) and coefficients of determination ( $R^2$ ) for surface temperatures of vegetation versus values from MSL, COS Z and NDVI. \*) Not significant at the 0.001 level.**

	MSL	COS Z	NDVI
R	-0.606	0.640	*
$R^2$	0.367	0.410	*

These linear correlation coefficients show that at time of satellite overpass COS Z has the strongest association with surface beech forests temperatures. This fact can be explained with the obvious warming of vegetation surfaces caused after sunrise on illuminated slopes. It is very likely that at the time of satellite overpass (10:09 local time) the temperature difference between illuminated and shaded surfaces is near its daily maximum (Conese et al. 1989). On 29 July 1993, the relationship between MSL and the surface temperature of the beech forests was weaker with a negative correlation coefficient. Moreover, including MSL along with COS Z in the calculation of a bivariate correlation coefficient did explain 79% of the variance of the surface temperature of the beech forests (table 2). On the contrary, according to Smith and Choudhury (1991), the simple linear regression of the surface temperature of the beech forests with the corresponding remotely sensed vegetation index amount had no significant correlation coefficient. In fact, no strong inverse relationships between surface temperatures and vegetation cover density indexes are evident for homogeneous forests with complete ground cover. This relationships exist only for crop and pasture lands where fractional vegetation cover varies from bare soil trough to crops with complete ground cover giving a wide range of surface temperature and NDVI values.

**Table 2. - Bivariate and multivariate coefficients of determination ( $R^2$ ) for surface temperatures of vegetation versus values from MSL, COS Z and NDVI. All coefficients are significant at the 0.001 level.**

	<b>MSL COS Z</b>	<b>MSL NDVI</b>	<b>NDVI COS Z</b>	<b>MSL COS Z NDVI</b>
$R^2$	0.790	0.374	0.413	0.792

A weak relationship between NDVI and surface temperature of the beech forests is still evident in combination with the topographic variables. The multiple correlation coefficient of the surface temperature of the beech forests with the three independent variables MSL, COS Z and NDVI increase their explained variance by less than 1% with respect to the bivariate correlation coefficient of the topographic variables alone with surface temperature.

## CONCLUSIONS

This study has defined some of the problems encountered when attempting to use remote sensed data to estimate surface temperatures of vegetation in complex

vegetated mountainous regions. In irregular terrains the influence of topography can strongly affect the thermal levels of vegetation. The comprehension of the sources of this influence is thus determinant for a correct interpretation of remotely sensed thermal data.

Results of the correlation analysis on 29 July 1993 in mid-summer sub-aridity conditions reveal that at time of satellite overpass COS Z alone accounted for 41.0% of the variance in surface temperature of the beech forests. This results indicate that insolation angle was, under the conditions of this experiment, a more important factor controlling surfaces temperatures of the beech forests than altitude or vegetation cover density. Moreover, the combined effect of differential heating and altitude explains 79.0% of the variance of surface temperature of the beech forests of Simbruini Mountains. On the contrary, the hypothesized negative relationship between NDVI and surface temperature is not significant due to the limited range of surface temperatures and fractional vegetation cover density values of the beech forests of the study area.

However, remotely sensed data provide only an instantaneous assessment of thermal conditions and the use of remote sensed thermal data is just a first step to describe surface temperatures of vegetation. Future works towards accurately define thermal trends of vegetation, should necessarily include field measurements to convert instantaneous satellite-derived thermal data to daily predictions.

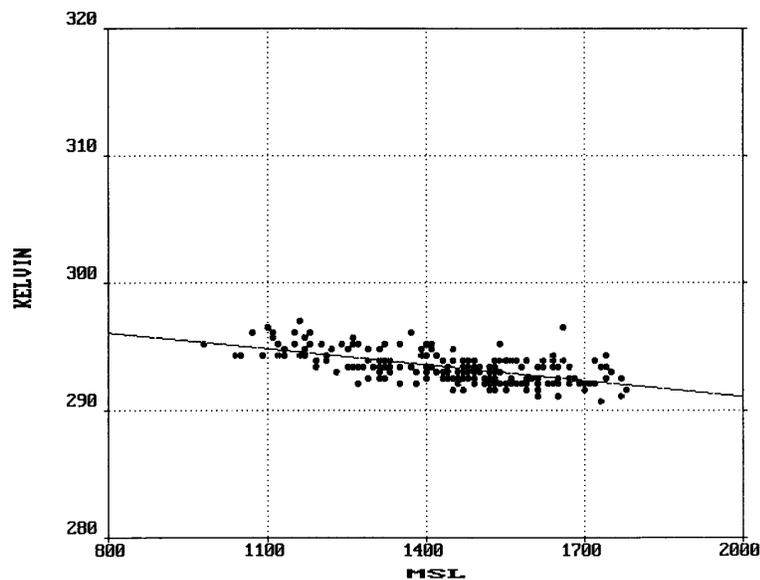


Figure 1. - Scattergram of surface temperatures versus MSL

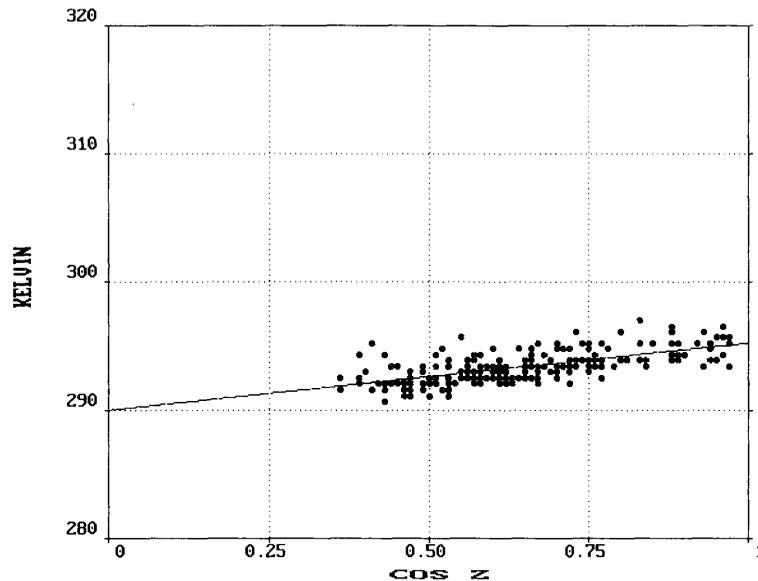


Figure 2. - Scattergram of surface temperatures versus COS Z

## REFERENCES

- Asrar, G., Harris, T.R., Lapitan, R.L. and Cooper, D.I. 1988. Radiative surface temperature of the burned and unburned areas in a tallgrass prairie, *Remote Sensing of Environment*, 24/447-457.
- Bartolucci, L.A., Chang, M., Anuta, P.E. and Graves, M.R. 1988. Atmospheric effects on Landsat TM thermal IR data, *IEEE Transactions on Geoscience and Remote Sensing*, 26/171-176.
- Blasi, C. 1994. *Fitoclimatologia del Lazio*. Regione Lazio, Ufficio Parchi e Riserve Naturali, Roma.
- Chavez, P.S. 1988. An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data, *Remote Sensing of Environment*, 24/459-479.
- Conese, C., Maselli, F. and Maracchi, G. 1989. Influence of orography on the thermal response of coniferous forests measured by TM data. User contribution to satellite remote sensing programmes, *Proceedings of the IX EARSEL Symposium held in Espoo, Finland, on 27 June 1989*, pp. 267-273.
- Congalton, R.G. 1988. A comparison of sampling schemes used in generating error matrices for assessing the accuracy of maps generated from remotely sensed data, *Photogrammetric Engineering and Remote Sensing*, 54/593-600.

- Friedl, M.A. and Davis, F.W. 1994. Sources of variation in radiometric surface temperature over a tallgrass prairie, *Remote sensing of Environment*, 48/1-17.
- Hope, A.S. and McDowell, T.P. 1992. The relationship between surface temperature and a spectral vegetation index of a tallgrass prairie: effects of burning and other landscape controls, *International Journal of Remote Sensing*, 13/2849-2863.
- Lathrop, R.G. and Lillesand, T.M. 1987. Calibration of Thematic Mapper thermal data for water surface temperature mapping: case study on the Great Lakes, *Remote Sensing of Environment*, 22/297-307.
- Markaham, B.L. and Barker, J.L. 1986. Landsat MSS and TM post-calibration dynamic ranges, exoatmospheric reflectances and at-satellite temperatures, *Landsat Technical Notes*, 1/3-8.
- Singh, S.M. 1988. Brightness temperature algorithms for Landsat Thematic Mapper data, *Remote Sensing of Environment*, 24/509-512.
- Smith, R.C.G. and Choudhury, B.J. 1990. On the correlation of indices of vegetation and surface temperature over south-eastern Australia, *International Journal of Remote Sensing*, 11/2113-2120.
- Smith, R.C.G. and Choudhury, B.J. 1991. Analysis of normalized difference and surface temperature observations over southeastern Australia, *International Journal of Remote Sensing*, 12/2021-2044.

### **BIOGRAPHICAL SKETCH**

Carlo Ricotta is a Ph.D. candidate in Botanical Sciences at the Department of Plant Biology of the University of Rome "La Sapienza", Italy, where he graduated in Geology in 1992. His research interests include the extraction of spatial information from remotely sensed imagery to study vegetation dynamic and landscape change.

Giancarlo Avena is Full Professor of Phytogeography and Director of the Department of Plant Biology at the University of Rome "La Sapienza". He graduated in Geology from the University of Rome "La Sapienza" in 1962 and specialized in Physical Geography and Geobotany. His current research interests include the application of remote sensing and associate techniques for the study of vegetation.

Fernando Ferri is technical co-ordinator of the Remote Sensing Laboratory of the Department of Plant Biology at the University of Rome "La Sapienza", where he graduated in Geology in 1981. Mr. Ferri provides technical consultation on GIS and remote sensing projects that affect vegetation mapping.

# Application of Non-Parametric Kernel Regression and Nearest-Neighbor Regression for Generalizing Sample Tree Information

Annika Kangas<sup>1</sup> and Kari T. Korhonen<sup>2</sup>

---

**Abstract** - Usually in a forest survey, large part of the measured trees are tally trees from which only elementary characteristics are measured. Part of the trees are sample trees, which are measured more thoroughly. To be able to utilize tally trees efficiently in the calculations the information available for sample trees has to be generalized for tally trees. The generalized information should be unbiased for areas of arbitrary size (or arbitrary groups of data). Also, the variation between sample plots and within sample plots should be realistic, if this data are used as a basis for simulations of forest development. These requirements may often be contradictory. In this paper applications of non-parametric kernel regression and nearest-neighbor regression for generalizing sample tree information are discussed. These methods may provide a satisfactory compromise between these requirements.

---

## INTRODUCTION

Two- or multi-phase sampling is applied in most forest inventory systems. The first phase sample (tally trees) consists of a large number of trees for which diameter and other easily measurable characteristics are measured. The second phase sample (sample trees) consists of trees measured more thoroughly. Height, age and additional diameters are the most typical characteristics measured for the sample trees. A third phase sample may be collected to derive volumes or biomass from sample tree characteristics. (Cunia 1986).

If two phase sampling is applied, the sample tree information has to be generalized for tally trees. This means that for every tally tree an expected value

---

<sup>1</sup> Research Scientist, Finnish Forest Research Institute, Kannus Research Station, P.O. BOX 44, FIN-60101 Kannus, Finland

<sup>2</sup> Research Scientist, Finnish Forest Research Institute, Joensuu Research Station, P.O. BOX 68, FIN-80101 Joensuu, Finland

of each sample tree characteristic with respect to measured characteristics is given. Most methods used are based on regression techniques (Korhonen 1993, Korhonen 1992, Cunia 1986, Kilkki 1979). The advantage of using regression models is that unbiased estimates for the population parameters are easily obtained.

The unbiased estimate of mean does not suffice, for example, if the predicted values are used as a data base for simulating the future development of forests (Ranneby and Svensson 1990). It is especially important to retain the natural variation, if the models used for predicting future development are non-linear (Moeur and Stage 1995). In this case, the different tree characteristics should harmonize, the between-plot and within-plot variation of the predictions should be as realistic as possible, the treewise and standwise results should be unbiased and, in addition, the mean in the population should be unbiased.

These requirements are often contradictory. It is difficult to formulate the model so that it is precise enough, provides realistic image of forests and gives unbiased estimates for arbitrary groups. A method, which is optimal in one respect may be unacceptable in other respects. What is acceptable, depends on the situation.

One example of a problem of this kind is describing the effect of geographic location on stem form. In the paper of Korhonen (1993) it was demonstrated that, in Finland, the stem form of Scots pine (*Pinus sylvestris*) depends on the geographic location besides the tree and stand characteristics. The traditional solution for this problem is to estimate separate models for areas of interest. In this approach, however, the effect of location cannot be taken into account for areas smaller or larger than the areas pre-defined and the effect of location is not continuous. Another solution is to take the location as a regressor to the applied model. A quadratic trend surface, for example, can be used for describing the large scale variation (Korhonen 1993). With the kriging method also the small-scale correlations can be taken into account (Ripley 1981). Unfortunately, the kriging method may be impractical for large data sets.

Often the models for generalizing different sample tree characteristics are estimated separately. Thus, it is possible to obtain values for volume and volume increment, which result in unrealistic estimate for volume increment percentage. Standard estimation procedures minimize componentwise losses, but what is needed is an ensemble of parameters (Louis 1984). Solution for this situation may be a system of simultaneous equations.

Non-parametric models can offer flexible solutions for generalizing sample tree data for different kind of situations. With non-parametric methods the effect of location can be taken into account with a simple procedure. It is also easy to generalize all the sample tree characteristics at the same time. Thus, it is easier to retain the covariance structure between the different tree characteristics. In this paper an application of semiparametric kernel regression and nearest-neighbor regression for generalizing sample tree information are discussed. The goal of this

paper is to compare the presented non-parametric methods from theoretical point of view and also with an empirical example.

## METHODS

### Kernel regression approach

A non-parametric estimate of a variable  $y$  at a point  $i$  is the weighted average of the measured values of  $y$ . The weight of a sample point depends on the differences in the values of the independent variables between the point of interest and the sample points. In this study a non-parametric regression model (Nadaraya 1964)

$$\hat{y}(x_j) = \frac{\sum_{i=1}^n y_i K\left[\frac{(x_j - x_i)}{h}\right]}{\sum_{i=1}^n K\left[\frac{(x_j - x_i)}{h}\right]} \quad (1)$$

was used, where  $y$  is the dependent variable,  $x_j$  ( $x_i$ ) is a vector containing the values of the independent variables at point  $j$  ( $i$ ),  $h$  is the window-parameter and  $K$  is the kernel function. In this study the kernel function used was a multivariate normal density function (Silverman 1986)

$$K(x_j) = (2\pi)^{-\frac{d}{2}} \exp\left(-\frac{1}{2} x_j' x_j\right) \quad (2)$$

where  $d$  is the dimension of the distribution (number of elements in vector  $x_j$  in 1).

The use of non-parametric kernel regression in the case of several independent variables is, however, difficult. When the number of independent variables increases, the data set may be surprisingly sparsely distributed in a high-dimensional Euclidean space. Thus, the applicable window-parameter values are too large to describe the relationships between the dependent and independent variables properly. Further, the model becomes difficult to interpret and impossible to demonstrate in a graphic form. (Härdle 1989).

Several methods for overcoming this problem have been presented, for example, by considering the linear combinations of the independent variables (see Härdle 1989, Moeur and Stage 1995). One obvious solution for the problem of several independent variables is to use semiparametric methods, i.e. combination of parametric and non-parametric methods. In this study, the residuals of a parametric model were smoothed with non-parametric kernel regression with coordinates as independent variables, in order to obtain localized estimates.

## Nearest-neighbor approach

A nearest neighbor estimator is a (weighted) average of  $k$  nearest neighbors of point  $j$ .

$$\hat{y}(x_j) = \frac{\sum_{i=1}^k y_i W(x_j, x_i)}{\sum_{i=1}^k W(x_j, x_i)} \quad (3)$$

In this study the weight function used was

$$W(x_i, x_j) = \frac{1}{1 + \sqrt{\sum_{m=1}^p a_m (x_{im} - x_{jm})^2}} \quad (4)$$

where  $a$  and  $k$  are parameters. The nearest neighbors are defined by some distance measure, in this study the nearest neighbors are those with largest weights. The parameter  $a_m$  defines the relative importance of the independent variable  $x_m$ . Another parameter of the nearest neighbor estimator (Eq. 3) is the number of neighbors included,  $k$ .

The kernel method and nearest-neighbor method are closely related methods. In nearest-neighbor approach the size of the neighborhood may vary, whereas in non-parametric kernel regression the size of the neighborhood is fixed and the number of neighbors varies. Nearest neighbor method is equivalent to kernel method with varying window width. The dimensionality problem in the nearest neighbor approach does not seem to be as serious problem as in the kernel method, because of the varying window width.

The methods differ also with other respects. For example, the nearest-neighbor produces a slightly rougher curve (Härdle 1989). This is due to discontinuity of the function. As the window moves, new observations enter to the  $k$  neighbors. If the new neighbor differs from the current average, there will be an abrupt change in the value of nearest-neighbor regression. Also, due to the fixed number of neighbors nearest neighbor method can be expected to be more biased near the boundaries than the kernel method.

Optimal values of the parameters  $a_m$  can be searched with cross validation method (Altman 1992). In this method each observation is predicted with data excluding the observation itself. Estimator with lowest mean of squared residuals is regarded best. In this study, the parameters used in this study are from Korhonen and Kangas (1995).

## AN EXAMPLE

### Material

The data used in this study were the pine sample trees of 8<sup>th</sup> NFI of eastern Finland. Only trees growing on site class II were included in the data. The data consist of 2063 pines measured on 375 plots.

Diameter at breast height and height were measured for the sample trees used in this study. Stem volumes were calculated using measured dimensions and volume functions of Laasasenaho (1982). For each plot several variables describing the site and growing stock were registered in the NFI data. These variables include location, altitude, site class, basal area of growing stock, dominant tree species, mean diameter and age of growing stock etc.

### Results

The nearest neighbor estimator was compared with the parametric and semiparametric estimator. The volumes of sample trees of NFI8 were estimated with other trees in the same data set. Mean and standard deviation of residuals of volume estimates were calculated using parametric estimator (Eq. 5), nearest neighbor estimator (Eqs. 3, 4) and semiparametric estimator (Eqs. 1, 2) with different values of parameters. The within-plot and between-plot variance components of volume were estimated to test how the different methods retain the initial variation in the data.

As the parametric estimator and the parametric part of the semiparametric model a function

$$v_{ij}/d_{ij}^2 = b_0 + b_1 d_{ij} + b_2 d_{ij}^2 + b_3 \ln(G_j) + b_4 T_j + b_5 d_{mj} + e_{ij} \quad (5)$$

was used. In Eq. 5  $v_{ij}$  is the volume of tree  $i$  on plot  $j$  ( $\text{dm}^3$ ),  $d_{ij}$  is the diameter at breast height (cm),  $\ln(G_j)$  is the natural logarithm of the basal area of the growing stock of the plot  $j$  ( $\text{m}^2/\text{ha}$ ),  $T_j$  is the mean age and  $d_{mj}$  is the mean diameter in the plot  $j$ . The only independent variables in the non-parametric part of the model were the coordinates. For the nearest neighbor approach the independent variables were  $d_{ij}^2$ ,  $G_j$ ,  $T_j$ ,  $d_{mj}$  and the coordinates. The  $b$ -parameters for parametric model and weights  $a_m$  for nearest neighbor approach are presented in table 1.

**Table 1. The variables in parametric model and their coefficients and, the variables in the nearest-neighbor regression and their weights.**

Variable	Coefficient	Variable	Weight
int	-0.0977	d <sup>2</sup>	0.227
d	0.0274	G	2.020
d <sup>2</sup>	-0.00035	T	1.812
ln(G)	0.0576	d <sub>m</sub>	1.937
T	0.00102	x	0.149
d <sub>m</sub>	0.00792	y	0.053

The results were calculated with several different window widths  $h$  for semiparametric approach and with different number of neighbors  $k$  for nearest neighbor approach. The results are presented in tables 2 and 3.

**Table 2. Mean and standard error of residuals of volume and between-plot and within-plot standard deviations of volume predictions with different window widths for semiparametric approach. The between-plot and within-plot standard deviations of true volumes are 254.1 and 179.6, respectively.**

Window width	Mean of residuals	Std. error	Std(plot)	Std(tree)
2 km	-2.19	49.78	249.3	181.5
4 km	-1.52	52.66	248.6	181.3
8 km	-0.99	57.67	248.8	181.3
16 km	-1.02	63.36	250.8	181.1
parametric	0.0	66.36	248.5	181.5

**Table 3. Mean and standard error of residuals of volume and between-plot and within-plot standard deviations of volume predictions with different number of neighbors for nearest-neighbor approach. The between-plot and within-plot standard deviations of true volumes are 254.1 and 179.6, respectively.**

Number of neighbors	Mean of residuals	Std. error	Std(plot)	Std(tree)
1	0.26	77.38	252.6	175.9
5	-0.54	65.96	248.1	171.2
10	0.058	65.40	246.7	170.2
15	0.52	65.84	245.1	169.4

The parametric estimates of volume are unbiased in the whole area, whereas

the non-parametric estimates generally are not unbiased. The smallest standard errors are obtained with semiparametric approach, using quite small window widths. The largest standard errors were obtained with nearest-neighbor method.

On the other hand, with parametric regression estimator the between-plot variation decreases and within-plot variation increases, when compared to the true variations. The semiparametric approach has no effect at all on the variance components. The most realistic variation is obtained with nearest-neighbor method with only one neighbor. From these results it can be concluded that the preferable method for generalizing sample tree information depends on the situation.

## DISCUSSION

In the standard estimation methods, like regression analysis, the primary goal is to obtain as accurate estimates as possible for individual observations. Thus, variance or MSE of individual predictions is minimized. In forest inventory, however, the goal is to obtain unbiased estimates of several parameters for arbitrary groups of observations. Consequently, minimizing the variance or MSE is not enough. The other criterias may be even more important than the MSE of individual observations. Also, the different requirements that are set to the results of generalization method may be contradictory.

In this study, in addition to the variance and bias also within-plot and between-plot variation of the predictions were considered. Other components, such as subgroup biases and variances could also have been considered. In this study the different components were not combined in order to find an optimal method, but it would be possible to define a "utility function" of the different components and their relative weights.

With the semiparametric approach the estimates of sample tree characteristics can be localized in order to obtain better estimates for subareas (-groups) of data (Kangas and Korhonen 1995). This is also true for nearest neighbor method. But in addition, with the nearest neighbor method the initial structure of the data can be retained fairly well (see also Korhonen and Kangas 1995). This is due to the fact that all the sample tree characteristics can be generalized at the same time. With a purely non-parametric kernel regression method this would also be possible. This approach, however, requires special attention to dimensionality problem, for example smoothing in one dimension with a linear combination of variables.

## REFERENCES

- Altman, N.S. 1992. An Introduction to Kernel and Nearest-neighbor Nonparametric Regression. *The American Statistician* 46:175-185.
- Cunia, T. 1986. Error of forest inventory estimates: its main components. In: *Estimating Tree Biomass Regressions and Their Error*. Proceedings of the

- Workshop on Tree Biomass Regression Functions and their Contribution to the Error of Forest Inventory Estimates. May 26-30, 1986. Syracuse, New York. pp. 1-13.
- Nadaraya, E.A. 1964. On Estimating Regression. Theory of Probability Application. No. 9:141-142.
- Härdle, W. 1989. Applied nonparametric regression. Cambridge University Press. 323 pp.
- Kangas, A. and Korhonen, K.T. 1995. Generalizing sample tree information with semiparametric and parametric models. *Silva fennica* 29(2):151-158.
- Kilikki, P. 1979. An outline for a data processing system in forest mensuration. *Silva Fennica* 13(4):368-384.
- Korhonen, K.T. 1992. Calibration of upper diameter models in large scale forest inventory. *Silva Fennica* 26(4):231-239.
- Korhonen, K.T. 1993. Mixed estimation in calibration of volume functions of Scots pine. *Silva Fennica* 27(4):269-276.
- Korhonen, K.T. and Kangas, A. 1995. Application of nearest-neighbor regression for generalizing sample tree information. Manuscript, 16 p.
- Laasasenaho, J. 1982. Taper curve and volume functions for pine, spruce and birch. *Communicationes Instituti Forestalis Fenniae* 108. 74 pp.
- Louis, T.A. 1984. Estimating a population of parameter values using Bayes and empirical Bayes methods. *J. Am. Stat. Ass.* 79(386):393-398.
- Moeur, M. and Stage, A.R. 1995. A most similar neighbor: an improved sampling inference procedure for natural resource planning. *For. Sci* 41(2):337-359.
- Ranneby, B. and Svensson, S.A. 1991. From sample tree data to images of tree populations. In: *Forest inventories in Europe with special reference to statistical methods. Proceedings of the International IUFRO S. 4.02 and S. 604 Symposium, May 14-16, 1990. Swiss Federal Institute for Forest, Snow and Landscape Research. Birmensdorf, Switzerland.*
- Silverman, B.W. 1986. *Density Estimation for Statistics and Data Analysis.* London. Chapman & Hall.

### **BIOGRAPHICAL SKETCH**

Annika S. Kangas is research scientist, Finnish Forest Research Institute, Kannus Research Station, Finland. She holds a D.Sc. in Forestry from University of Joensuu. Annika deals with forest inventory methods, especially with model based sampling techniques.

Kari T. Korhonen is research scientist, Finnish Forest Research Institute, Joensuu Research Station, Finland. He holds a D.Sc. in Forestry from University of Joensuu. Kari deals with forest inventory methods, especially with generalizing sample tree information.

# A Method for Assessing the Prediction Quality of Mechanistic Forest Growth Models

Biing T. Guan<sup>1</sup>, George Gertner<sup>2</sup> and Pablo Parysow<sup>3</sup>

---

**Abstract.** A method for assessing the prediction quality of mechanistic forest growth models was presented. The method consisted of four steps: assuming distributions for parameter values, parameter screening, outlining model behavior through sampling, and approximating model behavior based on the sampled points. The proposed method was then applied to a carbon balance stand level forest growth model. For the example, a Monte Carlo method was used to perform the sampling, and then a "patterned" artificial neural network was used for the approximation. It was found that the predictions of the model were relatively unbiased with respect to the initial parameter variances, and the variances of the predictions were mainly contributed by only a few parameters. Such information allows the model to be more closely examined and provides a way to improve the model.

---

## INTRODUCTION

For the past ten years, an important change in forest growth modeling has been the shift in modeling approach from empirical modeling towards mechanistic (or process) modeling; a shift that reflects the model developers' desire to try to gain better insights into how stands (or trees) grow and how future environment changes may affect forest ecosystems (Bossel, 1991). Though still not having the desired predictive power, mechanistic forest growth models are being developed and used more frequently to answer some serious questions regarding the likely impact of global environmental changes on various aspects of forest ecosystems. An example is the use of TREGRO, a mechanistic tree growth model, to assess the impact of O<sub>3</sub> on seedling growth of red spruce (Laurence *et al.*, 1993). Such a usage of mechanistic models raises an important question: what is the prediction quality (i.e., the amount of prediction bias and variance) of a mechanistic model

---

<sup>1</sup> Associate Professor, Department of Forestry, National Taiwan University, Taipei, Taiwan, Republic of China

<sup>2</sup> Professor, Department of Natural Resources and Environmental Sciences, University of Illinois at Urbana-Champaign, Urbana, IL

<sup>3</sup> Ph.D. graduate student, Department of Natural Resources and Environmental Sciences, University of Illinois at Urbana-Champaign, Urbana, IL

and how can such information be obtained? Existing prediction quality analysis methods, such as a Monte Carlo approach or an error propagation approach (e.g., Gertner, 1987), cannot handle mechanistic models effectively.

In this study, a method is proposed for assessing the prediction quality of mechanistic growth models. The framework of the method and layout of its key steps will be described first. Then, as an example, the method is applied to a mechanistic growth model. Finally, some comments will be made with respect to the generality of the proposed method.

## **GENERAL FRAMEWORK OF THE METHOD**

The core idea of the proposed prediction quality analysis method is as follows: we believe that the sampling intensity required by a Monte Carlo approach can be reduced through the use of a good approximation procedure while maintaining a specified accuracy of the quality analysis. We should emphasize at this point that the proposed method assumes that the target model is structurally and behaviorally valid (see Bossel, 1992) for the intended purposes. The proposed method has the following four key steps:

**Assuming Ranges and Distributions:** For the proposed method, we assume each parameter in the model has a fixed range and a distribution within the specified range. Unlike empirical models, parameters in mechanistic models are usually not estimated from data (some even argue that parameters in mechanistic models should not be estimated, e.g., Bossel, 1992); they are usually derived from the literature, small scale experiments, studies of related species, and from subjective sources when information is sparse. Though often viewed as exact, parameter values in mechanistic models should in reality be considered as interval estimates, that is, the “true” parameter value may be within a certain range, and the given parameter value is the most likely a value within the range.

**Screening Parameters:** Parameter screening is necessary for reducing the actual number of parameters that need to be examined. It is likely that in a mechanistic model not all parameters contribute equally to the prediction quality (bias and variance) of that model. Some parameters may have significant effects on prediction quality, some may have slight effects, and still others may have no effects. If those parameters that contribute little to prediction quality are screened out, then computational resources can be saved in the subsequent steps.

**Determining Monte Carlo Sampling Intensity:** For the proposed method, two sampling intensities need to be considered. The first intensity is the total number of points (each point represents a combination of variabilities of the parameters

and the initial values of the state variables) that will be used in the subsequent approximation. The sampled points roughly define the model's global behavior in the sampled space. The second intensity is the number of runs that need to be carried out at each sampled point to obtain actual prediction quality information at that sampled point (i.e., the local behavior around the sampled point).

**Selecting an Approximation Procedure:** Together with the previous step, this step determines the success of the proposed method. Choices abound for selecting an approximation procedure. One can choose from simple polynomials to complex nonlinear procedures, such as a spline or a neural network model. Different approximation procedures offer different advantages.

### EXAMPLE

The model that we used to demonstrate the proposed method is a mechanistic forest growth model developed by Valentine (1988). The model is a carbon balance model based on the pipe theory (Shinozaki *et al.*, 1964 a,b) and the self-thinning rule. The model is a stand-level growth model for even-aged stands. It has three state variables (basal area per unit area, active pipe length per unit area, and total woody volume per unit area) and 20 parameters. In this example, the pipe model was adapted for red pine (*Pinus resinosa* Ait.). The model was used to simulate the growth of red pine for 30 years, from age 30 to age 60. For this example, we will view the means as the "true" parameter values. Therefore, the final projections based on those means will be treated as the "unbiased" projections. Bias will thus be defined as the differences between the projected and the "unbiased" values.

**Step 1:** For the example, the mean value and the lower and the upper bounds for each of the parameters, as well as for the initial values of the state variables, were directly taken from Parysow (1994). It is assumed that each parameter, as well as the initial value of a state variable, has an unimodal and symmetric beta distribution bounded between the prescribed lower and upper bounds. The requirements on the shape of our beta distributions constrains the maximum variance that our beta distributions can have. For a beta distribution to be symmetric and unimodal, its two shape parameters need to be equal, and the values must be greater than 1. In this example, it was decided that the minimum value for the shape parameters of a beta distribution would be 1.5 which gives a beta distribution with a well-defined mode. Under these constraints, the standard deviation of a generic beta distribution in our analysis cannot be greater than 0.25; in other words, the maximum standard deviation of a parameter distribution is 25 percent of its distribution range.

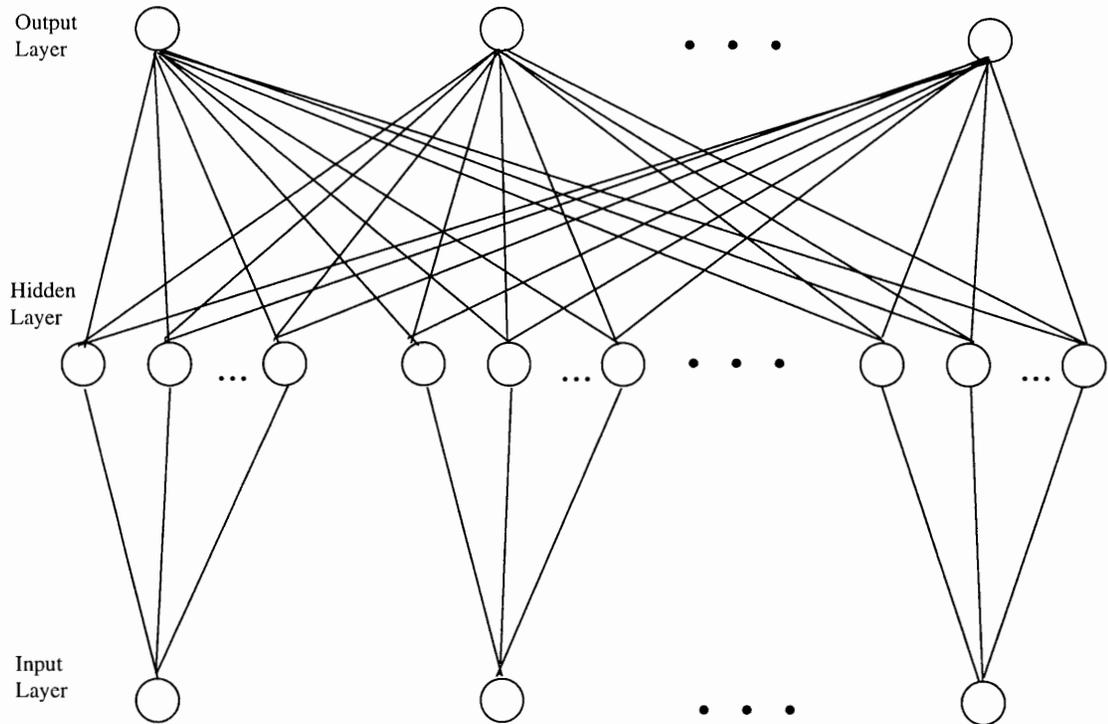
**Step 2:** Though the target model was a small model in terms of the number of parameters and state variables, it was still desirable to screen out parameters that had little influence on projection quality. The procedure used is Mann's test for a trend (Lehmann, 1975, pp. 290-297). Parameters were screened independently, and the state variables were not included in the screening procedure. The  $\alpha$ -level to include was set to be 0.2 since the intention was to screen out the parameters with little influences on the final projection quality.

**Step 3:** In this example, we chose a random sampling scheme to generate the necessary sample points as well as the local behavior at each sampled point. Two sets of data were generated. For the first set, the calibration data set, the sampling intensity was 64K for the first level sampling and 16K for the second level sampling. For the second set, the validation data set, the sampling intensities were 128K and 32K, respectively, for the first and second levels of sampling. It should be noted that although parameter screening was performed (Step 2), a complete sample was still done because the approximation results were used to validate the screening results.

**Step 4:** The approximation procedure used in this example was a feedforward, multilayer type of artificial neural network model. Artificial neural networks have been proven both theoretically (Cybenko, 1989; Hornik *et al.*, 1989) and empirically (e.g., Lapedes and Farber, 1987) to be excellent function approximators under general conditions. Artificial neural networks have also been used to approximate chaotic functions with good results (e.g., Lapedes and Farber, 1987). Though chaotic behavior might not exist in this example, it may exist in other mechanistic models, especially the complex ones, since most of the differential equations in existing mechanistic models have not been examined thoroughly for their behaviors. For these reasons, artificial neural networks have been chosen as the approximation tools in this example.

The networks used in this example all had an input layer, an output layer and a hidden layer. The networks used were different from the typical feedforward networks in two aspects. First, the nodes between the input layer and the hidden layer were not fully connected, and were without the bias connections (Figure 1).

The nodes in the hidden layer were divided into groups, and each group was connected only to one input node (groups were disjoint). Thus, each input was trained by a subset of hidden nodes. The activation function for the hidden nodes was the arc tangent ( $\tan^{-1}$ ) function. The output activation function was just a linear function (i.e., no squashing).



**Figure 1. Network architecture for the proposed network.**

Such a network architecture will produce outputs that can be partitioned according to the contribution of each input. The connection pattern of the network will isolate the contribution of each input, while the arc tangent activation function will make some hidden nodes inactive (i.e., set to 0's) when the corresponding inputs are inactive. The contributions from those inactive inputs to the overall network outputs will thus be 0's. To be able to partition the outputs according to the contribution of each input is important to the proposed procedure since it allows us to examine the importance of each input and to develop an error budget for the model.

The network we used in this example had five hidden nodes for each input node. The training algorithm used in this case study was a modified version of a random optimization procedure (Baba, 1989). The error function of the training method was the squared error function (i.e., the squared difference between the target and the actual outputs). Since this particular example involves a large number of training and validation data, and the selected training algorithm is parallel in nature, network training as well as validation were conducted on a

CM-2 parallel computer. Detail discussion of the training algorithm can be found in Guan *et al.* (1993).

## RESULTS AND COMMENTS

Since the results from the sampling showed that initial parameter variability did not affect the projection accuracy (i.e., very little bias) of the pipe model, we will focus on how initial parameter uncertainty affects the projection variances of the pipe model.

**Parameter Screening:** The results from parameter screening confirm our initial hypothesis that not all the parameters contribute significantly toward the projection variance of the pipe model. Table 1 lists the parameters that the initial screening procedure deemed to be significant (with p value, or calculated tail probability, smaller than 0.2) to overall projection variance for each of the three state variables. The entries are listed in a descending order according to the calculated p value of each parameter.

**Approximation Results:** Table 2 gives the validation results of the trained neural network for approximating the final projection variances of the three state variables. The mean error rates are within 4 percent for all the three state variables, and 99 percent of the approximations are within 14 percent of the target values. It should be noted that the trained network was for approximating the variances of the three state variables simultaneously (i.e., the network had three output nodes).

**Ranking the Importance of Parameters:** Since the neural network model we developed has the ability to isolate the variance contribution of an individual parameter, we can use this feature to rank the importance of individual parameters. Table 3 is an example of such a ranking. In the table, we can see that given the same amount of relative initial variability, the one with the largest contribution is the parameter S, a parameter that represents the effects of environmental conditions on tree growth. What is interesting is that the first four parameters in the table account for almost 75 percent of the total variability, and the first ten parameters account for about 95 percent of the total projection variance. The partitions of projection variances of the other two state variables also show similar patterns, though with different parameter ranking. Common to the projection variances of the three state variables is the relative importance of the parameter S. Given the same amount of relative initial parameter variability, this parameter accounts for roughly 37 percent, 63 percent and 55 percent of the final projection variances for the state variables basal area, pipe length and total

**Table 1. Results of parameter screening for the final projection variance for the three state variables. Entries are listed in a descending order according to the calculated p value of each parameter.**

Pipe Length	State Variable	
	Basal Area	Total Volume
S	S	S
Z	C	Z
Bp	Amax	Amax
P	Z	P
$\theta$	Bp	Bp
Tf	P	C
$\lambda$	$\theta$	Tf
Tz	Tf	V1
Pf	$\psi$	Tz
$\alpha$	$\lambda$	$\theta$
Amax	F	$\psi$
Bf	Tz	F
		$\alpha$

**Table 2. Validation results for the trained neural network for predicting the final projection variance of the three state variables. Numbers represents the relative error rates (in percentage)**

	State Variables		
	Pipe Length	Basal Area	Total Volume
Mean	3.79	1.38	2.03
Minimum	0.00	0.00	0.00
25 Quantile	1.70	0.52	0.76
Medium	3.38	1.12	1.61
75 Quantile	5.24	1.96	2.82
95 Quantile	8.79	3.52	5.36
99 Quantile	13.05	4.92	7.83
Maximum	30.04	11.68	17.23

volume, respectively. Such a ranking in parameter contribution provides us with clues to further improve the performance and projection quality of a model.

As the contribution of an individual parameter toward the overall prediction quality becomes available, we can use such information to develop error budgets for the model, to test hypotheses, and to analyze the statistical power of the tests, steps that are all necessary in using mechanistic models to make informed decisions.

**Table 3. An example of a partition of the final projection variance of total basal area per unit area based on the contribution of each parameter. The initial standard deviation for each parameter was set at 5% of the parameter's range.**

<b>Parameter</b>	<b>Contribution</b>	<b>Relative Importance (%)</b>	<b>Cumulative Importance (%)</b>
S	2.70	37.52	37.52
C	1.30	18.07	55.59
Basal Area	0.98	13.54	69.14
Z	0.48	6.63	75.77
Amax	0.42	5.82	81.59
Pipe Length	0.29	3.98	85.57
Bp	0.25	3.46	89.03
P	0.18	2.52	91.54
$\lambda$	0.11	1.57	93.11
$\theta$	0.09	1.24	94.35
Total Volume	0.09	1.22	95.57
Tf	0.07	1.04	96.61
$\psi$	0.06	0.85	97.45
V2	0.06	0.83	98.29
F	0.03	0.42	98.71
Tz	0.02	0.35	99.05
Rz	0.02	0.30	99.35
$\alpha$	0.02	0.27	99.62
Rp	0.00	0.14	99.76
Bf	0.00	0.12	99.88
Rf	0.00	0.08	99.97
V1	0.00	0.02	99.99
Bz	0.00	0.01	100.00
<b>Total</b>	<b>7.20</b>	<b>100.00</b>	

Comparing the rankings in Tables 1 and 3, not including the state variables, we can see that the parameter screening procedure we proposed can indeed screen out the parameters that have little influence on the final projection variance of total basal area. The proposed screening procedure also was effective in screening out the parameters that had little effect with respect to the other two state variables. With a conservative screening criterion, we believe that the proposed screening procedure can effectively reduce the number of parameters that need to be considered while still maintaining the overall approximation accuracy. When the

model of interest is large and complex, parameter screening may save a large amount of computation resources in the future.

As mentioned earlier, many approximation procedures can be used for the proposed method, and one such method is orthogonal response surface analysis. For the same pipe model, we have also developed a second-order orthogonal response surface model based on a fractional factorial design (Gertner *et al.*, 1996). The orthogonal response model and the neural network model yielded almost identical results, though the former requires less computation since it was based on a fractional factorial design. However, when the model of interest has a large number of parameters, or when the higher order approximation is called for (e.g., the approximation is highly non-linear), then it is our belief that a neural network approach may be more appropriate. A major difference between the two models lies in the way the total projection variance (or bias) is partitioned. For an orthogonal model, we may not be able to isolate the contribution of an individual parameter since interaction with other parameters may be significant. Our neural network model, on the other hand, has the ability to isolate the individual parameter contribution, and the partition is exact (though not necessary orthogonal).

The method proposed in this study is also general in the sense that the method can be used as effectively in analyzing large and complex models as in small and simple models, though complex or large models may require more computational resources. For example, for complex multi-component models, we can adopt a bottom-up, divide-and-conquer approach, that is, we analyze the behavior of individual components, and then analyze the behavior of components toward the overall model. Thus, it is our belief that the proposed method will contribute significantly toward a more rational use of mechanistic models in ecological research.

#### **ACKNOWLEDGMENTS**

The authors thanked the National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, for granting the necessary computer time on the CM-2 computer.

## REFERENCES

- Bossel, H., 1991. Modeling forest dynamics: Moving from description to explanation. *Forest Ecology and Management* 42: 129-143.
- Bossel, H., 1992. Real-structure process description as the basis of understanding ecosystems and their development. *Ecological Modelling* 63: 261-276.
- Cybenko, G., 1989. Approximation by superposition of a sigmoidal function. *Mathematics of Controls, Signals and Systems* 2: 303-314.
- Gertner, G., 1987. Approximating precision in simulation projections: an efficient alternative to Monte Carlo methods. *Forest Science* 33: 230-239.
- Gertner, G., P. Parysow and B. T. Guan. 1996. Projection variance partitioning of a conceptual forest growth model with orthogonal polynomial. To appear in *Forest Science*.
- Guan, B. T. and G. Z. Gertner, and D. Kowalski. 1993. Modeling training site vegetation coverage probability with a random optimization procedure: an artificial neural network approach. *In: the Proc. of the Conference of Applications of Artificial Neural Networks IV, Orlando, Florida*, pp. 682-688.
- Hornik, K., Stinchcombe, M., and White, H., 1989. Multilayer feedforward networks are universals approximators. *Neural Network* 2: 359-366.
- Lapedes, A., and Farber, R., 1987. Nonlinear signal processing using neural networks: Prediction and system modelling. Technical Report LA-UR-87-2662, Los Alamos National Laboratory, New Mexico.
- Laurence, J. A., Kohut, R. J., and Amundson, R. G., 1993. Use of TREGRO to simulate the effects of ozone on the growth of red spruce seedlings. *Forest Science* 39(3): 453-464.
- Lehmann, E. L., 1975. *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco. 457 p.
- Parysow, P. F., 1994. A procedure to approximate prediction uncertainty in ecological conceptual models. Unpublished M.S. Thesis, University of Illinois at Urbana-Champaign, Urbana, Illinois. 126 p.
- Shinozaki, K., Yoda, K., Hozumi, K. and Kira, T., 1964a. A quantitative analysis of plant form - the pipe model theory. I. Basic analysis. *Japanese Journal of Forest Ecology* 14: 97-105.
- Shinozaki, K., Yoda, K., Hozumi, K., and Kira, T., 1964b. A quantitative analysis of plant form - the pipe model theory. II. Further evidence of the theory and its application in forest ecology. *Japanese Journal of Forest Ecology* 14: 133-139.
- Valentine, H., 1988. A carbon balance model of stand growth: a derivation employing pipe-model theory and the self-thinning rule. *Annals of Botany* 62:389-396.

# Statistical Analysis of Error Propagation in National Level Carbon Budgets

C. J. Cieszewski<sup>1</sup>, D. P. Turner<sup>2</sup>, and D. L. Phillips<sup>3</sup>

---

## Abstract

<sup>4</sup> Existing national-level carbon budget studies typically do not conduct formal uncertainty analyses. In cases where attempts at addressing carbon budget uncertainties have been made, they are generally limited to informal variations of sensitivity studies. The reasons for limited effort towards uncertainty analysis in carbon budget studies include: 1) unknown or poor statistical properties of the input variables; 2) use of arbitrary relationships and analytically intractable discrete relationships in carbon budget models; 3) a high level of aggregation and complexity in calculations; and at times, 4) the interactive nature of the developed carbon budget models which does not lend itself to the application of a Monte Carlo approach. Development of a generic uncertainty analysis framework for carbon budget studies is proposed here. The key component will be a flexible error propagation program designed for operation within Monte Carlo simulations and using statistical software. The simulation software will have an open architecture, such that it can be implemented at a user defined level of aggregation with regard to land cover type in a variety of countries. Error estimates will be developed for variables including total biomass carbon, net ecosystem production, and carbon emissions associated with land use change. The error propagation program will include a library of functions and difference equations useful for replacing discrete relationships such as growth and yield tables, and providing a basis for analytical treatment of potential error propagation.

---

<sup>1</sup>Forest Biometrician, ManTech Environmental Research Services Corp., Corvallis, OR.

<sup>2</sup>Forest Ecologist, ManTech Environmental Research Services Corp., Corvallis, OR.

<sup>3</sup>Research Biologist, U.S. EPA National Health & Environmental Effects Laboratory, Corvallis, OR.

<sup>4</sup>We acknowledge Sandra Brown, U.S. EPA National Health & Environmental Effects Laboratory, Corvallis, OR, for the initiation of this study.

## INTRODUCTION

The Framework Convention for Climate Change (FCCC) calls for national-level inventories of net greenhouse gas emissions, including carbon (CO<sub>2</sub>-C) sources and sinks associated with forests and land-cover change. The International Panel on Climate Change (IPCC) has subsequently worked with the Organization for Economic Cooperation and Development (OECD) to produce an initial set of protocols for estimating the land base flux (IPCC 1995). A number of national-level studies have been made, e.g., Makundi *et al.* 1992, Kauppi *et al.* (1992), Kolchugina and Vinson (1993), Kurz *et al.* (1992, 1994), Maclaren and Wakelin (1991), Turner *et al.* (1993, 1995), but none of them conducts formal error analyses and the only attempts at addressing carbon budget uncertainties are limited to informal versions of sensitivity analyses, e.g., Kurz *et al.* (1994).

There are several reasons why so little progress has been made in the issue of uncertainty analysis in the carbon budget studies. One reason is the limited information about properties of errors in input variables used in the estimation of greenhouse gas flux (Robinson 1989). It is impossible to estimate statistical properties of propagated errors if the statistical properties of the input errors are not known and a formal sensitivity analysis might be the only possibility to properly quantify the uncertainties in carbon budget models.

Furthermore, when there is a fair amount of information about the errors associated with the input variables, their statistical properties may not lend themselves to parametric error propagation analysis. For example, forestry data are usually characterized by large variation, strong correlation between different variables, and frequently a lack of normality in their distributions. The properties of distributions of errors associated with such variables can be analytically intractable within even moderately complex systems; parametric statistical analysis rely on normally distributed variables, and even just a correlation between analyzed variables dramatically increases the relative complexity of the involved calculations.

Another problem for analytical error analysis is the use of arbitrary non-functional relationships in computations. Such systems in principle are analytically intractable. A propagation of errors can be analytically traced only within continuous mathematically defined systems of equations on a limited level of complexity. With large and complex systems it is usually more practical and at times the only possibility to apply simulation methods and sensitivity analysis. However, practical applications of simulation methods require that models be written in forms that lend themselves to large numbers of repetitions in the model executions and easily changeable input selections.

To solve these problems and to address the concerns of uncertainty analysis on the carbon budget and flux computation a Monte Carlo simulation approach is suggested. This paper addresses some concerns related to using the Monte Carlo simulation approach and describes its application with regards to national level carbon budgets and national level carbon flux estimation.

## TYPES AND SOURCES OF ERRORS

The types of errors are in general divided into absolute errors and relative errors. The errors can be systematic or random, or random with a constant

bias. In the carbon budget modeling and flux estimation the errors can originate from the data collection, i.e., inventory and data compilation, and from the carbon budget modeling framework.

There are three major sources of errors (Cunia 1985, Gertner and Köhl 1992) associated with the inventory: measurement errors, sampling errors, and regression errors. Sampling errors can be the greatest contribution to total error contained in inventory data. Exact estimation of inventory errors is difficult since the actual exact quantities measured in forestry are never well known and definable. Estimation of errors for inventory data can be based on combination of analytical methods associated with sampling procedures and simulation techniques applied to data compilation calculations.

Most errors generated/propagated within carbon budget modeling framework are subject to such sources as:

- Errors in Given Input Data.
- Round-off Errors During the computations.
- Truncation and Discretization Errors.
- Errors Resulting from Simplifications in the Mathematical Model.
- “Human” Errors and Machine Errors.

### **ASSUMPTIONS ON CARBON BUDGET DESIGNS**

Since it is desirable to have one consistent computational schema defining all potentially applicable carbon budgets, the framework developed here will be applicable only to a lowest common denominator of carbon budget elements. The lowest common denominator is assumed to be the lowest available inventory of carbon quantities. This means that even in the most information rich carbon budgets, e.g., the US carbon budget, the analysis will assume availabilities of error distributions of carbon quantities entering the carbon budget model on the lowest level of resolution, e.g., live tree carbon yield (TCY). The error distributions arising from all lower levels of derivation will have to be analyzed outside of this generalized framework.

In the example of the TCY (in the US carbon budget), its quantity is not measured but derived from timber inventory information that is based on measurements of diameters, heights, ages, densities, species compositions, etc., with associated sampling errors, measurement errors and other errors. Those errors are propagated by applied formulas and assumptions that are used to derive TCY, and the result associates certain error distributions that have to be derived or estimated outside of the generic carbon budget uncertainty analysis framework. Exact derivation of the error distributions for TCYs would depend on designs and assumptions of each individual inventory and models applied to the measured quantities. For the purpose of this analysis these error distributions have to be based on expert opinion.

The carbon budget uncertainty analyses are considered here as two potential studies. The first study is the uncertainty analysis of the carbon pools at different aggregation levels, i.e., analysis of uncertainty associated with a carbon inventory at one point in time that summarizes a specified number of

carbon pools. The second study is the uncertainty analysis of the carbon flux estimation, i.e, estimation of uncertainties associated with estimated changes in the carbon pools over time.

A third potential area of effort could concentrate on analysis of uncertainty associated with estimation of a periodic carbon flux based on two given carbon inventories conducted at different times. This approach would involve analysis of conditional uncertainties depending on a knowledge of statistical distributions of uncertainties associated with each of the two available inventories. These would likely be correlated and potentially would not have normally distributed errors. Nevertheless, this problem could also be addressed through applications of Monte Carlo simulations, though the results would not be easily generalized due to the conditional aspect of such approach.

### Definition of the Carbon Pools and their Budget

A basic entity of any carbon budget is a carbon store at a given level of aggregation. This store is defined by its level of aggregation, denomination, area, and yield, i.e., quantity per unit area. An amount of carbon in a store at a given aggregation level and for a given denomination is calculated for this store as a product of the area and yield of this store. If at the lowest level of aggregation in detailed models a store consists of sub-units of carbon storage within one basic unit area, i.e., basic carbon pools having defined individual yields, the store yield is then simply a product of the store area and its respective sum of pertaining pool yields. All of the carbon stores are parts of corresponding higher aggregation stores and they are obtained from adding up sums of lower aggregation stores of the same category. Carbon stores can be represented within the analysis framework on an arbitrary level of aggregation. An example of a carbon store at high level of aggregation is tropical forests in Amazonia. An example of a carbon store on a low level of aggregation is live tree carbon (TCY) in a monospecific stand under specific management practices in one management unit and at one age.

A carbon store on a specific aggregation level  $l$  is always defined as simply a sum of carbon stores within the same category,  $m$ , at  $l - 1$  level of aggregation, i.e.,  $C_{(l)m} = \sum_1^{N_{(l-1)m}} A_{(l-1)m} Y_{(l-1)m}$  where:  $C_{(l)m}$  is carbon storage at the aggregation level ( $l$ ) in the category  $m$ ;  $N_{(l-1)m}$  is number of carbon stores, or pools, at level ( $l - 1$ ) in the category  $m$ ;  $A_{(l-1)m}$  is area of carbon store  $m$  at  $l - 1$  aggregation level; and  $Y_{(l-1)m}$  is the yield of carbon store  $m$  at aggregation level  $l - 1$ .

The rule described by this equations is applied recursively to the carbon stores. If the areas or yields on a level ( $l - 1$ ) are not defined they are computed from level ( $l - 2$ ). If the carbon quantities at level ( $l - 2$ ) are not defined they are computed from level ( $l - 3$ ), and so on, until a level  $k$  at which the quantities of carbon are defined as the basic model input. At that point the summations can be brought up gradually from level  $k$  to level ( $l$ ). Thus any carbon budget is defined as a sum of products between areas and their yields from the lowest aggregation level of carbon pools, even though, for stores composed of sums of basic carbon pools within the store area  $C_{(l)m} = \sum_{i=1}^{N_{(l-1)m}} \sum_{j=1}^{n_{(l-1)m}} A_{(l-1)m_i} Y_{(l-1)m_{ij}}$ , where:  $n$  is the number of basic pools in the

lowest aggregation level store; and  $Y_{(l-1),m_j}$  is a yield for pool  $j$  in category  $m$  at level  $(l-1)$ .

### Carbon Budget Marginal Uncertainties

First order marginal uncertainties around the carbon budget are defined here by the partial derivatives of the sum of basic carbon pools within the store area with respect to its input variables  $A$  and  $Y$ , i.e.,  $\frac{\partial C_{(l)m}}{\partial A_{(l),m_i^*}} = \sum_{j=1}^{n_{(l-1),m}} Y_{(l-1),m_i^*j}$  and  $\frac{\partial C_{(l)m}}{\partial Y_{(l-1),m_i^*j^*}} = A_{(l-1),m_i^*}$ .

It is clear from the first order marginal uncertainties that the second order of marginal sensitivities in accordance with the general carbon budget scheme defined here would be equal to zero for both area and stocking related errors. The non-dimensional quantities in the first order linear marginal uncertainties are defined as relative values:

$$\frac{\partial C_{(l)m}/C_{(l)m}}{\partial A_{(l),m_i^*}/A_{(l),m_i^*}} = \frac{A_{(l-1),m_i^*} \sum_{j=1}^{n_{(l-1),m}} Y_{(l-1),m_i^*j}}{\sum_{i=1}^{N_{(l-1),m}} \sum_{j=1}^{n_{(l-1),m}} A_{(l-1),m_i} Y_{(l-1),m_{ij}}} \quad (1)$$

and

$$\frac{\partial C_{(l)m}/C_{(l)m}}{\partial Y_{(l-1),m_i^*j^*}/Y_{(l-1),m_i^*j^*}} = \frac{Y_{(l-1),m_i^*j^*} A_{(l-1),m_i^*}}{\sum_{i=1}^{N_{(l-1),m}} \sum_{j=1}^{n_{(l-1),m}} A_{(l-1),m_i} Y_{(l-1),m_{ij}}} \quad (2)$$

The total error of the carbon budget of category  $m$  at level  $l$  is defined as  $\sum_{j=1}^{n_{(l-1),m}} Y_{(l-1),m_i^*j} + A_{(l-1),m_i^*}$ . If we define the total error in the carbon budget at  $l$  level of aggregation and  $m$  category as  $\Delta C_{(l)m}$  and the errors associated with inputs of the components as  $\Delta A_{(l-1),m_i}$  and  $\Delta Y_{(l-1),m_{ij}}$ , then:

$$\Delta C_{(l)m} = \sum_{i=1}^{N_{(l-1),m}} \left( A_{(l-1),m_i} \sum_{j=1}^{n_{(l-1),m}} \Delta Y_{(l-1),m_{ij}} + \Delta A_{(l-1),m_i} \sum_{j=1}^{n_{(l-1),m}} Y_{(l-1),m_{ij}} \right) \quad (3)$$

In relative terms the total error of the carbon budget of category  $m$  at level  $l$  is defined as:

$$\frac{\Delta C_{(l)m}}{C_{(l)m}} = \frac{\sum_{i=1}^{N_{(l-1),m}} \left( A_{(l-1),m_i} \sum_{j=1}^{n_{(l-1),m}} \Delta Y_{(l-1),m_{ij}} + \Delta A_{(l-1),m_i} \sum_{j=1}^{n_{(l-1),m}} Y_{(l-1),m_{ij}} \right)}{\sum_{i=1}^{N_{(l-1),m}} \sum_{j=1}^{n_{(l-1),m}} A_{(l-1),m_i} Y_{(l-1),m_{ij}}} \quad (4)$$

Using the marginal uncertainties for uncertainty analysis based on formal sensitivity analysis is the most effective way to treat the problem of calculating the uncertainties in the responses of carbon budgets simulations due to uncertainties of the input parameters. This method is effective and relatively simple for implementation and it can be used with changing input error assumptions. It is also suitable for identifying most important sources of errors.

## ERROR ANALYSIS OF CARBON FLUX ESTIMATION

It is assumed here that the carbon flux is computed from simulations of changes in carbon budget over time rather than from differences between carbon budgets at different times. In such a dynamic approach a simulation of error propagation in carbon flux computations should be responsive to initial input values and errors associated with them as well as to potential deviations from any underlined growth patterns. Since none of the responses can be readily simulated using yield tables the simulation model should be based only on continuous functions.

To facilitate the error propagation of the simulated system, functions simulating changes in carbon pools over time should be self-referencing. That is, they should be based on dynamic equations or difference equations that are driven by initial conditions of the independent and dependent variable together with the independent variable so that estimated inventory errors can be directly incorporated into a simulation inputs.

Development of the dynamic equations is necessary for those carbon budgets that are currently based on yield tables. An example here is the US forest C budget (Turner et al. 1995) which is based on a complete forest inventory, a set of stand level C budgets which specify the age-specific pools and flux associated with C in 5 pools (tree, understory, forest floor, soil and woody debris), and a specified forest harvest.

In this example, the model is based on static carbon yield tables. Therefore the first step in computation of the carbon flux error propagation is to develop a set of dynamic equations replacing the static yield tables. All the steps for adopting this framework to the Monte Carlo error analysis approach can be outlined as follows:

1. For each stand level C budget, replace the age-specific pools/tables and flux values by equations (e.g., tree C vs age).
2. Identify the important sources of error within the algorithms that are used to create the stand level C budgets.
3. Determine an appropriate formulation of the error distribution for each case (either from the literature or using an assumption).
4. Create the mechanism within the simulation package to vary the matrix of error estimates for successive Monte Carlo runs.
5. Run the pools and NEP flux calculation multiple times using the varying values for specific pools, ratios, area, fluxes, etc.
6. In the case of the transition flux, determine an appropriate error distribution for the areas involved and perform the Monte Carlo simulation.
7. Examine Monte Carlo output for specific carbon pools and fluxes, for different levels of aggregation, and for the whole system.

Only the first step from the above list has been addressed here. Development of equations to replace tables is necessary to be able to consistently

propagate errors, examine the influence of changes in simulated trends on the error propagation, and streamlining the computational procedures that would be involved in numerous repetitions of all the computations involved in Monte Carlo simulations. Furthermore, to examine the potential impact of inventory errors on the carbon budget simulations it is advantageous to form the yield equations as self-referencing functions or dynamic equations. Such equations have been developed for all the US carbon budget yield tables and are presented in the following sections of this paper.

### **Example of Flux Estimation: US Carbon Budget**

#### **The Tables**

The carbon yield over age data used in the US carbon budget consisted of tables with carbon yields of five basic categories of carbon pools: i) live tree carbon; ii) soil carbon; iii) forest floor carbon; iv) understory carbon; and v) coarse woody debris carbon. Included were tables for both yields and yearly increments for all these categories. The tables contained data for total carbon yields and growth increments for total carbon, i.e., sums of the five categories, and for merchantable wood volume.

Based on species associations, the tables were aggregated into 50 groups and plotted for visual inspections and preliminary analysis of potential growth patterns. Some tables showed drastic inconsistencies in growth patterns suggesting that either the grouping criteria were too general or the tables are deficient. This means that the future work should address the problem of growth pattern inconsistencies within species grouping in the US carbon yield tables and more detailed work will have to be done to define more consistent clustering of the carbon yield patterns.

Ranges in stand age vary from 0–100 to 0–180 years for different species, management units, and management practices. Some tables include before and after thinning values. There are about 500 sets of tables for about 50 combinations of species. After separating the tables with before and after thinning values there are over 6000 tables describing the carbon changes as a function of age, in a > 1 Megabyte file with a total of about 200,000 discrete numbers, that would have to be interpolated for any ages not included in the tables.

Given that every table for each single carbon pool contains about 40 numbers that may need further interpolations, developing a set of equations with a unique set of coefficients for each table would be a great advance in streamlining the calculations involved in the US carbon budget. The equations would not only reduce the number of involved references, but also would allow for more accurate and continuous computations of carbon quantities over time. Furthermore, dynamic equations, unlike static tables, or for that matter even static equations, allow for implementation of error propagation analysis in flux calculations and for clustering of the tables into common growth patterns even at variable productivity levels between different tables.

## The Dynamic Equations

Five types of dynamic equations have been developed for the US Carbon Budget Flux estimation for the five different pools. Nonlinear regression fittings of these equations varied in performance and the number of needed sets of coefficients to describe all the relevant tables. The best performance was achieved with the equation for forest floor carbon; only one set of coefficients was required to fit well all of the forest floor tables.

The developed dynamic equations are truly base-age invariant and they are based on different base equations and different assumptions of the cross-sectional changes between different tables. The analytical derivation details of all equations are beyond the scope of this paper and are described in another manuscript. Graphs, coefficients and statistical summaries of the relevant regression analysis for the calibration of the dynamic equations are available on request from the principal author of this paper. The forms of the developed dynamic equations are described below under subsequent headings denoting the five relevant carbon pools of the US Carbon Budget. The following symbols are used in the illustrated below dynamic equations: i)  $\alpha, \beta, \dots$  are the equation estimable coefficients; ii)  $t$  is the age of prediction; iii)  $t_0$  the age of reference at which a value of carbon is known and used as the equation input; iv)  $y_0$  is the known value of carbon at age  $t_0$  that is used as the equation initial value; and v)  $Y$  is the carbon yield at age  $t$ , given the initial value of carbon at age  $t_0$  equal to  $y_0$ .

**Forest Floor Carbon** — The growth and yield of carbon in forest floor data was the most consistent for all tables and the developed model was able to describe the carbon growth and yield for all tables ( $R^2 > 99.93\%$ ) with a single set of coefficients. The curves generated by this equation have different values at origin and they are polymorphic with variable asymptotes. The equation has the following form:

$$Y(t) = \frac{t^\alpha (\beta + t_0^\alpha) (\delta\gamma + y_0)}{(\beta + t^\alpha) (\delta\beta + t_0^\alpha + \delta t_0^\alpha)} + \frac{y_0 (1 + \beta/t_0^\alpha) - \gamma}{1 + \beta/t_0^\alpha + 1/\delta} \quad (5)$$

**Live Tree Carbon** — Fifty sets of coefficients were fitted to all the predefined groups of tree live carbon. The curves generated by this equation are equal to zero at origin and they are polymorphic with variable asymptotes. The consistency of growth patterns was fairly good with few minor exceptions and the fitted model explained much of the variation between different yield tables and it has the following form:

$$Y(t) = \frac{t^\alpha \left( \frac{\beta^2}{\gamma^\alpha} + y_0 + \sqrt{\frac{4\beta^2 y_0}{t_0^\alpha} + \left(-\frac{\beta^2}{\gamma^\alpha} + y_0\right)^2} \right)}{2t^\alpha + \frac{4\beta^2}{-\frac{\beta^2}{\gamma^\alpha} + y_0 + \sqrt{\frac{4\beta^2 y_0}{t_0^\alpha} + \left(-\frac{\beta^2}{\gamma^\alpha} + y_0\right)^2}}} \quad (6)$$

**Coarse Woody Debris Carbon** — Like the live tree carbon the coarse woody debris carbon growth and yield is described by one equation with 50

sets of coefficients. The fits are not as good as for the live tree carbon but were good enough to justify the grouping of the tables into common patterns described by the following equation:

$$Y = \alpha t - \delta t^3 + (y_0 - \alpha t_0 + \delta t_0^3) \frac{1 - \beta t + \gamma t^2}{1 - \beta t_0 + \gamma t_0^2} \quad (7)$$

**Understory Carbon** — The carbon dynamics of the understory were very inconsistent. The changes over time were not continuous and seemed to consist of arbitrary numbers. No generalizations could be developed within the time frame available for this analysis. The understory carbon dynamics equation had to be fitted to each table individually resulting in 647 sets of coefficients. Even with individually fitted coefficients the regressions did not explain much of the variation seemingly present in this variable. All 647 sets of the coefficients relate to one equation of relatively low complexity  $Y(t) = \frac{t(1-\beta t+\gamma t^2)y_0}{t_0-\beta t_0^2+\gamma t_0^2}$ .

**Soil Carbon** — The yields of soil carbon were constants in almost all tables and accordingly their increments were zero. In a few tables, the soil carbon was increasing with a constant growth rate. A simple difference equation with only one coefficient ( $Y(t) = y_0 \frac{1+\beta t}{1+\beta t_0}$ ) can describe this behaviour without a need for regression analysis.

### SHORT DISCUSSION AND CONCLUSION

The described study aimed towards developing a generic framework for error analysis on national level carbon budgets and flux tracking models. A number of steps have been outlined for this development and two important steps have been advanced. The outlined points of consideration for potential Monte Carlo applications may be helpful in planning and undertaking similar studies.

The synopsis of the formal sensitivity analysis through analytical definition of marginal uncertainties potentially can improve efforts of uncertainty analysis through sensitivity analysis in comparison to these attempted in the past, e.g., Kurz *et al.* (1994). The conceptualized generalization of a carbon budget modeling framework could be used to develop a generic carbon budget dynamic simulation software that would be used for both formalized sensitivity analysis and Monte Carlo simulations of different hypothetical carbon budgets.

Finally, the development of the dynamic equations for the US carbon budget illustrates potential gains in the development of such equations with regard to reducing the amount of information which must be stored and increasing the flexibility in model testing and applications. Development of these equations makes possible error propagation analysis in carbon flux calculations and model behaviour investigation that could not be attempted with tables.

### LITERATURE CITED

- Cunia, T. 1985. On volume tables and their contribution to the error of forest inventory volume estimates. P. 205–222 *in* Proc. National Forest Inventory in Europe Workshop. University of Freiburg–Faculty of Forestry, Freiburg, Federal Republic of Germany.

- Gertner, G., and M. Köhl. 1992. An assessment of some nonsampling errors in a national survey using an error budget. *Forest Science* 38: 525–538.
- Intergovernmental Panel on Climate Change. 1995. *IPCC Guidelines for National Greenhouse Gas Inventories*. 3 vols., IPCC, Bracknell, UK.
- Kauppi, P. E., K. Mieliköinen, and K. Kuusela. 1992. Biomass and carbon budget of European forests, 1971 to 1990. *Science* 256: 70–74.
- Kolchugina, T. P., and T. S. Vinson. 1993. Carbon sources and sinks in forest biomes of the Soviet Union. *Global Biogeochemical Cycles* 7: 291–304.
- Kurz, W. A., M. J. Apps, T. M. Webb, and P. J. McNamee. 1992. Informative report. The carbon budget of the Canadian forest sector: Phase I, NOR-X-326. Forestry Canada, Edmonton, Alberta, Canada.
- Kurz, W. A., M. J. Apps, S. J. Beukema, and T. Lekstrum. 1994. 20th century carbon budget of Canadian forests. *Tellus* 47 B: 170–177.
- Maclaren, J. P., and S. J. Wakelin. 1991. Forestry and forest products as a carbon sink in New Zealand. *FRI Bulletin*. Number 162. Ian Bryce, Rotorua, New Zealand.
- Makundi, W., J. Sathaye, and O. Masera. 1992. Carbon emission and sequestration in forest: Case studies from seven developing countries, Vol. 1: Summary. Report No. LBL-32199, Lawrence Berkeley Laboratory, University of California, Berkeley, CA.
- Robinson, J.M. 1989. On uncertainty in the computation of global emissions from biomass burning. *Climatic Change* 14:243–262.
- Turner, D. P., J. J. Lee, G. J. Koerper, and J. R. Barker, editors. 1993. The forest sector carbon budget of the United States: carbon pools and flux under alternative policy options, EPA/600/3-93/093. United States Environmental Protection Agency, Environmental Research Laboratory, Corvallis, Oregon, USA.
- Turner, D. P., G. J. Koerper, M. E. Harmon, and J. J. Lee. 1995. A carbon budget for forests of the conterminous United States. *Ecological Applications* 5(2): 421–436.

### BIOGRAPHICAL SKETCH

Chris J. Cieszewski is a Forest Biometrician with ManTech Environmental Research Services Corp., Corvallis, OR. He holds a Ph.D. in Forest Biometrics from University of Alberta, has been working in his field since 1986 and specialises in development of mixed-effects base-age invariant models.

David P. Turner is a Forest Ecologist with ManTech Environmental Research Services Corp., Corvallis, OR. He holds a Ph.D. in Botany from Washington State University and specialises in the area of forest nutrient cycling.

Don L. Phillips is a Research Biologist with the U.S. EPA National Health & Environmental Effects Laboratory, Corvallis, OR. He holds a Ph.D. in Biology from Utah State University, and has worked in EPA's climate change research program since 1988.

# Monte Carlo Simulations of Nonlinear Size-Age Relationships

Ronald E. McRoberts<sup>1</sup>

---

**Abstract.**-Monte Carlo simulations are an accepted and useful procedure for determining the effects of uncertainty in input variables and parameters on the uncertainty of derived output variables. Monte Carlo procedures are particularly useful for growth model applications when the model is nonlinear in the parameters. For example, a nonlinear growth model may be fit to variables derived from size-age data, and then Monte Carlo simulations may be used to estimate the parameter covariance matrix and a confidence interval resulting from repeatedly applying the growth model beginning with a set of initial conditions. However, when simulated values of size variables contain randomly generated components, negative growth may occur. An example using height-age data and the Weibull growth model is used to illustrate and compare approaches to this problem and to justify an assertion that it may be difficult to precisely predict future height.

---

## INTRODUCTION

Monte Carlo simulation techniques have been recognized as a valid method for determining the effects of uncertainty in input variables and parameters on the uncertainty of derived output variables (McRoberts 1996, 1995, 1994, Mowrer 1994, 1992, Makela 1988, Gertner and Dzialowy 1984). They are particularly useful for estimating parameter covariance structures and for propagating uncertainty through nonlinear growth models when linear approximations perform poorly, when the mathematical formulation is complex, and when predictor variables depend on predicted variables at a previous stage.

The underlying objective is to estimate the bias and uncertainty in size predictions obtained from repeatedly applying an annual growth model beginning with an observed initial condition. Calibration data for the growth model consist of a series of size-age observations for a typical tree whose future size is to be predicted. Nonlinear regression is used to estimate the model parameters, and Monte Carlo simulations are used to estimate the covariance structure of the parameter estimates and confidence intervals for future size predictions. Two difficulties arise when applying simulations to this problem: (1) accommodating the repeated measures nature of the data when simulating observations and when calibrating the growth model; and (2) dealing with negative growth that occurs

---

<sup>1</sup> *Mathematical Statistician, North Central Forest Experiment Station, 1992 Folwell Ave., St. Paul, MN 55108.*

when a simulated size observation is less than a simulated size observation at a previous age due to randomly generated residuals.

## DATA

The calibration data consist of either 13 or 14 height-age observations for each of six co-dominant yellow birch trees from the same forest stand. The data for this study are a subset of the yellow birch stem analysis data collected by Carmean (1978) and used to construct northern hardwoods site index curves. They were collected from a plot in a fully stocked, even-aged stand on the Chequamegon National Forest in northern Wisconsin. Annual rings were counted on disks cut at each section point. For Carmean's analyses, a height-age graph was constructed for each tree and was examined for signs of early suppression, top breakage, or dieback. Such disturbed trees were discarded as were those from even-aged stands whose age range for dominant and codominant trees exceeded 10 years.

These height-age data are characterized as longitudinal, repeated measures data: (1) they were obtained by repeatedly observing the ages of the same trees at different heights; (2) they were necessarily ordered in time and position and allowed no randomization; and (3) the trees from which they were obtained were assumed to constitute a representative sample from the population of dominant and co-dominant trees. The serial correlation exhibited by longitudinal, repeated measures data due to the lack of randomization must be accommodated in parameter estimation routines if unbiased confidence intervals are to be obtained.

## METHODS

### Height-age model

McRoberts (1996, 1995) has shown that the height-age relationship for these data is adequately described using a Weibull model,

$$Y_i = \beta_1 (1 - e^{-\beta_2 x_i^{\beta_3}}) + \epsilon_i \quad [1]$$

where  $Y$  is height,  $x$  is age, and  $\beta = (\beta_1, \beta_2, \beta_3)$  is a parameter vector to be estimated. The residuals,  $\epsilon$ , are distributed  $N(0, \sigma^2 \Lambda)$  with elements,  $\lambda_{ij}$ , of the correlation matrix,  $\Lambda$ , structured as

$$\lambda_{ij} = \rho^{|x_i - x_j|}, \quad [2]$$

where  $\rho$  is the annual serial correlation and  $\sigma^2$  is the residual variance.

To obtain the height-age relationship for a typical tree, population parameters for Eq. [1] were estimated using a nonlinear mixed effects modeling approach. Models representing the height-age relationship for a population of trees contain

both fixed effects (population parameters) and random effects (individual parameters) and are described as mixed effects models. To simultaneously accommodate the mixed effects and longitudinal, repeated measures nature of the data, the first-order linearization procedure for nonlinear mixed effects models described by Davidian and Giltinan (1995) was used. Two assumptions underlie this procedure: (1) the same model form, but with possibly different parameter values, is adequate to describe the height-age relationship for all trees; and (2) the same residual variance, correlation structure, and correlation parameter value apply for all trees. The procedure applies least squares procedures to estimate the fixed and random model parameters and maximum likelihood procedures to estimate the common residual variance,  $\sigma^2$ , and annual serial correlation,  $\rho$ .

The residuals for estimated individual tree height-age curves exhibited homogeneity of variance. This result is consistent with the assumption that the standard deviation of growth is proportional to the expectation of growth, because height increments between adjacent observations were always approximately four feet. An estimate,  $c_\sigma$ , of the proportion was obtained by a second application of the nonlinear mixed effects model routine using a covariance structure based on this assumption and an assumption of annual height growth correlation,  $\rho$ .

### Height growth model

Three variables were derived from the original height-age observations: (1) average growth, the difference in height between adjacent sections divided by the corresponding difference in age; (2) average height of adjacent sections; and (3) average age of adjacent sections. These variables were used to develop a model for predicting average annual growth as a function of average age and average height. Because the Weibull model has been shown to adequately describe the height-age relationship for these data, height growth was modeled using a differential form of Eq. [1],

$$Y_i = \alpha_1 x_{1i}^{\alpha_2 (x_{2i} - \alpha_3)} + \epsilon_i \quad [3]$$

where  $Y$  is average height growth,  $x_1$  is average age,  $x_2$  is average height, and  $\alpha = (\alpha_1, \alpha_2, \alpha_3)$  is a parameter vector to be estimated. The parameters,  $\alpha$ , of Eq. [3] correspond to the parameters,  $\beta$ , of Eq. [1] as follows:  $\alpha_1 = \beta_2 \beta_3$ ,  $\alpha_2 = \beta_3 - 1$ ,  $\alpha_3 = \beta_1$ . The residuals,  $\epsilon$ , are distributed  $N(0, V)$  where the elements of  $V$  are calculated on the basis of the following assumptions: (1) variance is proportional to the square of expected growth; (2) variance is inversely proportional to the number of years over which average growth is calculated; and (3) the correlation among observations is derived from Eq. [2].

### Confidence Intervals

A 6-step Monte Carlo procedure was used to estimate an approximate 95% confidence interval around the estimated height-age curve obtained from repeated,

annual applications of the growth model. To estimate these confidence intervals, random numbers were generated to simulate uncertainty from two sources: (1) variability in the estimates of model parameters, and (2) residual variability around the curves resulting from the parameters estimates.

### **Step 1. Selection of a typical tree**

Estimates of the population parameters,  $\beta$ , and the common, individual tree estimates of  $\sigma^2$ ,  $c_\sigma$ , and  $\rho$  were used to generate data for a typical tree.

### **Step 2. Simulation of height-age observations**

Four-foot increments, beginning at 8 feet and ending at 64 feet, were selected, and Eq. [1] was solved using the population parameter estimates selected in Step 1 to determine the corresponding expected ages. A multivariate normal vector of residuals,  $\epsilon$ , was randomly generated using the common values of  $\sigma^2$  and  $\rho$  and was added to the selected heights to simulate a height-age series. Random selection of residuals may cause simulated height at a subsequent age to be less than simulated height at a previous age. Three approaches to this negative height growth problem were implemented: (Approach A) ignore the problem; (Approach B) reject simulated height-age series exhibiting negative height growth; and (Approach C) reject simulated height-age series whenever the absolute value of a residual is greater than expected height growth. Average height growth, average age, and average height were calculated from these simulated data.

### **Step 3. Estimation of growth model parameters**

The parameters for Eq. [3] were estimated from the simulated average annual data calculated in Step 2 with weighted nonlinear least squares (WNLS). The values of  $c_\sigma$  and  $\rho$  selected in Step 1 were used to calculate a weight matrix that was constant for all simulations.

### **Step 4. Simulation of annual height predictions**

Separate 50-year series of annual height-age observations were simulated beginning with the first age calculated in Step 2, and beginning with ages in ten year increments thereafter. For each year in each 50-year series, height was calculated as the sum of previous height, predicted annual height growth, and a randomly generated residual. Predicted height was obtained from Eq. [3] using the parameters estimated in Step 3. The residuals were multivariate normal with serial correlation,  $\rho$ , and standard deviations calculated as the product of  $c_\sigma$  and predicted height growth. The simulated height-age series were saved for calculation of confidence intervals in Step 6. Additional values of  $c_\sigma$  ( $c_\sigma=0.50$ ,  $c_\sigma=0.60$ ) were also used to determine the effect of this parameter on bias and uncertainty. The same three approaches to negative height growth described in Step 2 were implemented in the simulation of these annual height-age observations.

### Step 5. Replication

Steps 2-4 were replicated 1000 times for each approach to the negative height growth problem and each value of  $c_\sigma$ .

### Step 6. Construction of confidence intervals

For each year in the 50-year series, the 1000 annual height predictions were ordered from smallest to largest, and those corresponding to the 2.5, 50.0 and 97.5 percentiles were selected. The curve resulting from the 50th percentile predictions across all 50 ages approximates a median curve, while the other two percentiles jointly approximate a 95% confidence interval. Finally, an expected height-age curve with no random variability added was calculated using Eq. [3] with parameters derived directly from the population parameters selected in Step 1.

Strictly for comparison purposes, 1000 simulations were also conducted using an ordinary nonlinear least squares (ONLS) approach. These simulations were similar to those previously described with the following exceptions: (1) independent, normal residuals with variance,  $\sigma^2$ , were used in Steps 2 and 4; (2) the negative height growth problem was completely ignored in Steps 2 and 4; and (3) ONLS was used for parameter estimation in Step 3.

## RESULTS AND CONCLUSIONS

Parameter estimates for the height-age population curve (Fig. 1) were obtained using the nonlinear mixed effects model routine:  $\beta=(82.0107,-0.0123,1.1409)$ ;  $\sigma^2=1.92$ ;  $c_\sigma=0.39$ ; and  $\rho=0.67$ . The cumulative distribution of standardized residuals indicated the assumption of normal residuals was justified.

For each simulation of 15 height-age observations in Step 2, there were approximately 0.15 rejections for Approach B and approximately 0.50 rejections for Approach C. In Step 4, rejections occurred approximately 0.20 times for each simulation of 50 annual observations for Approach B and approximately 0.50 times for Approach C. With respect to Approach A, confidence intervals for both Approaches B and C exhibited bias. For Approach B, both the median curve and lower confidence limit exhibited positive bias, while for Approach C the median curve was unbiased, but the confidence interval was too narrow. Although the magnitude of the bias was not pronounced for  $c_\sigma=0.39$ , it increased with larger values of  $c_\sigma$  (Fig. 2).

The median curve for the ONLS method was unbiased, but the 95% confidence interval was too large. This result confirms the necessity of accommodating serial correlation and heterogeneity of variance when using Monte Carlo simulations and when estimating parameters from longitudinal, repeated measures data.

The results suggest that for  $c_\sigma \geq 0.40$ , an unpleasant dilemma may arise: either (1) accept statistical bias as the price to pay for biological realism, or (2) tolerate biologically impossible conditions in order to avoid statistical bias. From a

strictly statistical perspective, however, negative height growth poses no inherent conceptual difficulties. Even though the expected curve is monotonically increasing, the underlying statistical assumptions do not require subsequent observations to be larger than previous observations; the only requirement is that the residuals satisfy the distributional assumptions. Thus, procedures for constructing confidence intervals that properly reflect residual uncertainty and uncertainty in parameter estimates must ignore negative height growth.

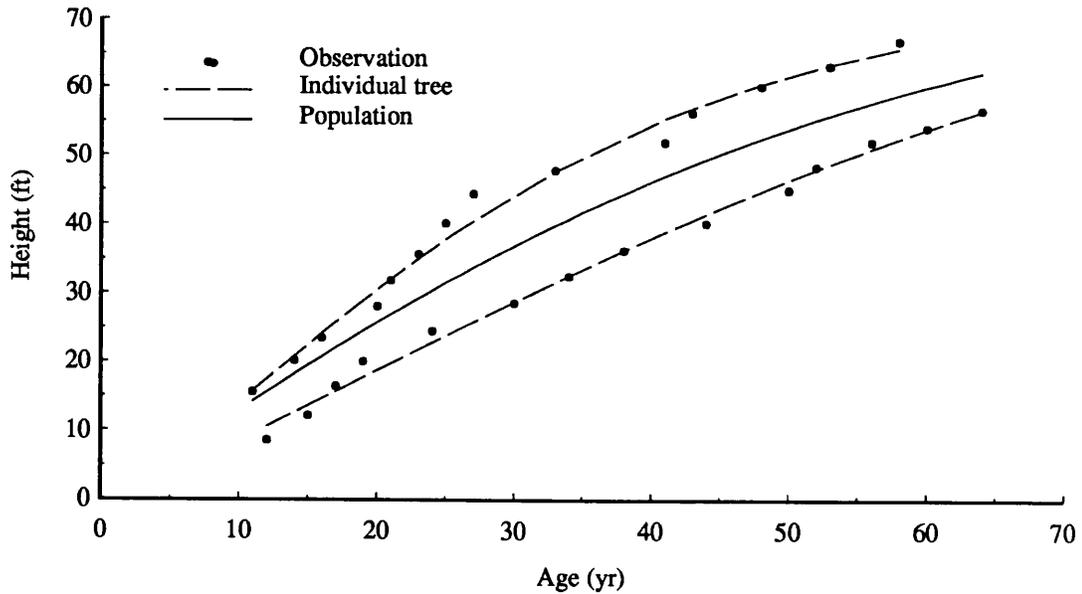


Figure 1. Height-age relationships.

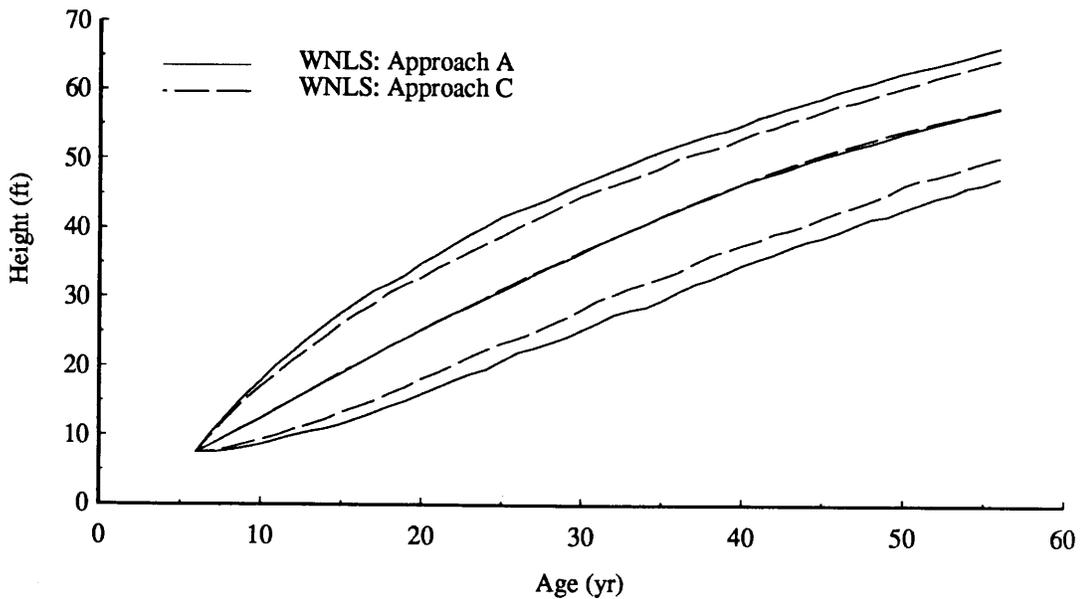


Figure 2. Median curves and 95% confidence intervals for  $c_p=0.50$ .

Overall, by the fiftieth year of predicted height, confidence intervals increased to a width of about 15-20 feet, depending on the initial condition (Fig. 3). Coefficients of variation were estimated by dividing  $\frac{1}{4}$  the width of the confidence interval by predicted height. The estimates were generally greater when the initial age and the length of the prediction interval were small (Table 1). The general guideline that coefficients of variation ought not exceed about 0.10 (Gertner 1994) suggests that for some situations it may be extremely difficult to precisely predict future height using an annual growth model. In addition, the results obtained in this study must be considered nearly optimal, because the calibration data represented only the tree to which the growth model was applied. For an actual application, the calibration data would include variation among a large sample of trees as an additional source of uncertainty and would produce much larger coefficients of variation.

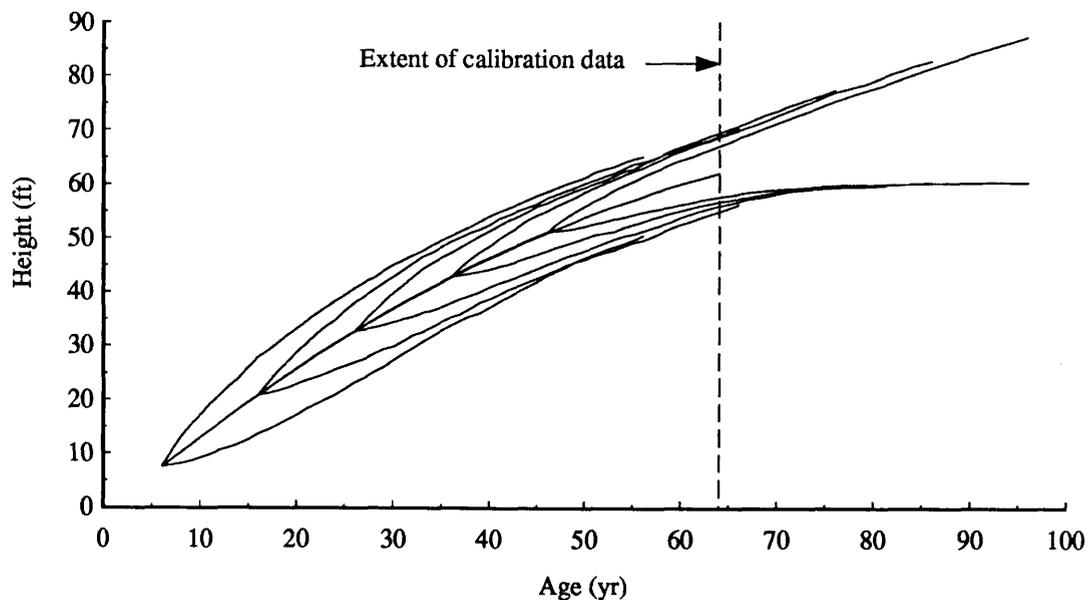


Figure 3. 95% confidence intervals.

Table 1. Coefficients of variation.

Initial age (yr)	Length of prediction (yr)				
	10	20	30	40	50
6	0.17	0.13	0.11	0.08	0.07
16	0.09	0.08	0.07	0.07	0.06*
26	0.05	0.05	0.06	0.05*	0.06*
36	0.04	0.05	0.05*	0.07*	0.08*
46	0.03	0.04*	0.06*	0.08*	0.09*

\* Beyond range of calibration data.

## REFERENCES

- Carmean, W.H. 1978. Site index curves for northern hardwoods in Northern Wisconsin and Upper Michigan. USDA For. Serv. Res. Pap. NC-160.
- Gertner, G.Z. 1994. A quality assessment of a Weibull based growth projection system. In: Growth and Yield from Estimation from Successive Forest Inventories. Proc. IUFRO Conf., Copenhagen, Denmark, June 14-17, 1993.
- Gertner, G.Z., and Dzialowy, P.J. 1984. Effects of measurement errors on an individual tree-based growth projection system. *Can. J. For. Res.* 14:311-316.
- Makela, A. 1988. Performance analysis of a process-based stand growth model using Monte Carlo techniques. *Scand. J. For. Res.* 3:315-331.
- McRoberts, R.E. 1996. Estimating variation in field crew estimates of site index. *Can. J. For. Res.* (in press).
- McRoberts, R.E. 1995. Assumptions underlying estimation methods for nonlinear mixed effects models. In: Am. Stat. Assn. 1995 Proc. Biom. Section, August 13-17, 1995, Orlando. FL. (in press).
- McRoberts, R.E. 1994. Variation in forest inventory field measurements. *Can. J. For. Res.* 24:1766-1770.
- Mowrer, H.T. 1994. Monte Carlo techniques for propagating uncertainty through simulation models and raster-based GIS. In: Proceedings of the International Symposium on the Spatial Accuracy of Natural Resource Data Bases, May 16-20, 1994, Williamsburg, VA, R.G.Congalton, ed. Am. Soc. Photogrammetry and Rem. Sens. pp179-188.
- Mower, H.T. 1992. Open-architecture Monte Carlo uncertainty analysis for forest stand dynamics Models. In: Research on Growth and Yield with Emphasis on Mixed Stands, IUFRO Centennial Meeting Session of Section 4.01, Berlin/Eberswalde, German, August 31-September 4, 1992. pp 75-84.

## BIOGRAPHICAL SKETCH

Ron McRoberts is a member of the Forest Inventory and Analysis unit of the North Central Forest Experiment Station, USDA Forest Service, where he develops individual tree growth models. His interests include nonlinear regression modeling, repeated measures, and propagation of variance.

# Cross-correlations among Single Tree Growth Models

Hubert Hasenauer<sup>1</sup>, Robert A. Monserud<sup>2</sup>,  
Timothy G. Gregoire<sup>3</sup>

---

**Abstract.** - Single tree growth and yield models basically consist of a number of equations to update tree parameters over time. Although it seems reasonable to assert that these equations are interrelated from a biological standpoint, it is customary to consider them independently and apply linear or nonlinear regression techniques separately rather than jointly. Using more than 7,500 Norway spruce (*Picea abies*) trees, we compare an individual tree basal area increment model, a height increment model, and a crown model using least square methods separately and jointly by applying three stage least square (3SLS) regression techniques. Results indicate a strong cross-equation correlation between the basal area and the height increment model and the height increment and the crown model. This suggests that the use of joint regression techniques would be superior.

---

## INTRODUCTION

Individual tree forest growth and yield models usually employ a set of equations to describe stand development over time. A typical single-tree stand simulator (Monserud 1975, Wykoff et al. 1982, Burkhart et al. 1987, Hasenauer 1994) consists of different equations for predicting periodic diameter or basal area increment, height increment, and the probability of mortality for each sample tree. These equations are usually developed separately.

From a biological standpoint, it seems reasonable to assert that the change in a tree's basal area, height, and risk of mortality are not uncorrelated phenomena (Dixon et al. 1990). Depending on their interrelationships, joint estimation of the models' parameters may be necessary in order to provide estimates that are consistent, or it may be desirable in order to provide more precise estimates than can be obtained otherwise.

Seminal work by Aitken (1934-35), Haavelmo (1943), Theil (1953),

---

<sup>1</sup> Biometrician, Institut für Waldwachstumsforschung, Universität für Bodenkultur, Wien, Austria.

<sup>2</sup> Biometrician, Intermountain Research Station, USDA Forest Service, Moscow, ID 83843, USA.

<sup>3</sup> Biometrician, Dept. Forestry, Virginia Polytechnic Institute & St. Univ., Blacksburg, VA, USA.

Zellner (1962), and Zellner and Theil (1962) resulted in almost all methods currently available for estimating the parameters in intercorrelated (simultaneous) systems of equations.

The objective of this paper is to evaluate and compare joint versus separate regression techniques for single tree growth and yield modeling. We estimate the parameters of individual tree models for basal area increment, height increment, and crown ratio using least squares methods separately and jointly by applying two-stage (Theil 1953) and three-stage least squares (Zellner and Theil 1962) techniques. We specifically investigate (1) the differences in the estimated coefficients and (2) the correlation between the predictions.

## METHODS

### Independent Regressions

We begin with a system of three individual tree growth equations for stand conditions in Austria: basal area increment, height increment, and crown ratio. These equations were developed independently using the same dataset.

**Basal area increment:** After eliminating the qualitative site descriptors chosen by Monserud and Sterba (1996), which reduced the variance explained only by 2.6%, we are left with the following model for  $Y_1$ :

$$\ln(\Delta BA) = a + b_1 \cdot \ln(D) + b_2 \cdot D^2 + b_3 \cdot \ln(C) + c_1 \cdot BAL + c_2 \cdot CCF^2 + d_1 \cdot ELEV^2 + d_2 \cdot SL^2 + e_1 \quad (1)$$

with  $\Delta BA$  the 5-year basal area increment (outside bark),  $D$  the diameter at breast height (1.3 m) in cm,  $C = (1/CR) - 1$  where  $CR$  is the crown ratio,  $BAL$  the basal area ( $m^2/ha$ ) of trees larger in diameter than the subject tree,  $CCF$  the crown competition factor of Krajicek et al. (1961),  $ELEV$  the elevation in hectometers, and  $SL$  the tangent of the slope angle ( $\%/100$ ).

**Height increment:** Hasenauer and Monserud (1996a) used a similar formulation to predict 5-year height increment  $\Delta H$ , where  $H$  is the tree height, and all other parameters as previously defined. The second equation ( $Y_2$ ) in the system is:

$$\ln(\Delta H) = a + b_1 \cdot \ln(D) + b_2 \cdot H^2 + b_3 \cdot \ln(C) + c_1 \cdot BAL + c_2 \cdot CCF + d_1 \cdot ELEV^2 + d_2 \cdot SL + e_2 \quad (2)$$

**Crown ratio:** To ensure that the predictions of crown ratio (defined as the crown length divided by tree height) are bounded between 0 and 1, Hasenauer and Monserud (1996b) chose a logistic function. After linearizing the logistic and rearranging, we are left with the following logarithmic transformation of crown ratio ( $Y_3 = \ln(C)$ ):

$$\ln(C) = a + b_1 \cdot H/D + b_2 \cdot H + b_3 \cdot D^2 + c_1 \cdot BAL + c_2 \cdot \ln(CCF) + d_1 \cdot ELEV + d_2 \cdot SL^2 + d_3 \cdot SL \cos AZ + d_4 \cdot SL \sin AZ + e_3 \quad (3)$$

where  $C = (1/CR) - 1$ ,  $H/D$  is the height/diameter ratio (m/cm),  $AZ$  is the azimuth in radians, and all other parameters are as previously defined. Change in crown ratio is not available, because height to crown base was not remeasured after the initial inventory.

## Simultaneous Equation Systems

In the system above, it is likely that the errors  $\epsilon$  in equations (1) - (3) are intercorrelated because they are associated with various attributes of the same tree. If this is the only common influence among the three equations, then Zellner's (1962) seemingly unrelated regression (SUR) procedure would be appropriate because the equations are related through contemporaneous correlations in the variance-covariance matrix. We write the multivariate regression model as

$$Y = X\beta + \epsilon \quad (4)$$

where  $Y$  is a  $3n \times 1$  vector of dependent (endogenous) variables,  $X$  is the  $3n \times (p_1 + p_2 + p_3)$  design matrix,  $\beta$  is the  $(p_1 + p_2 + p_3) \times 1$  vector of coefficients to be estimated, and  $\epsilon$  is the  $3n \times 1$  error vector. The errors  $\epsilon$  have fixed mean  $E[\epsilon] = 0$  and variance

$$V[\epsilon] = E[\epsilon\epsilon'] = \begin{pmatrix} W_1 & W_{12} & W_{13} \\ W_{21} & W_2 & W_{23} \\ W_{31} & W_{32} & W_3 \end{pmatrix} = \Omega \quad (5)$$

where  $W_i = \sigma_i^2 I$  are the main diagonal variances and  $W_{ij} = \sigma_{ij} I$  are the covariances, with  $I$  the  $n$ -dimensional identity matrix.

The ordinary least squares (OLS) estimator of  $\beta = (\beta'_1, \beta'_2, \beta'_3)$  is

$$b = (x'x)^{-1} x'y \quad (6)$$

with variance

$$\text{cov}(b) = (x'x)^{-1} (x'\Omega x) (x'x)^{-1} \quad (7)$$

In situations where  $\sigma_{ij} \neq 0$  and  $\sigma_{ij} \neq \sigma^2$  for some constant  $\sigma^2$ , then  $b$  is inefficient and the usual estimator of (7), namely

$$c\hat{\text{ov}}(b) = (x'x)^{-1} \hat{\sigma}^2 \quad (8)$$

is biased. Zellner's (1962) seemingly unrelated regression procedure is a form of feasible generalized least squares (Judge et al. 1980) in which  $\Omega$  is estimated by  $\hat{\Omega}$ , where  $\hat{\Omega}$  has the same form as (5) but with moment estimators, say  $\hat{\sigma}_i^2$  and  $\hat{\sigma}_{ij}$ , in place of the unknown parameters  $\sigma_i^2$  and  $\sigma_{ij}$ . With  $\hat{\Omega}$ , an asymptotically efficient estimator of  $\beta$  is

$$\hat{\beta} = (x'\hat{\Omega}^{-1}x)^{-1} x'\hat{\Omega}^{-1}y \quad (9)$$

The variance of  $\hat{\beta}$  is estimated by

$$\hat{\Sigma} = \text{cov}(\hat{\beta}) = (x'\hat{\Omega}^{-1}x)^{-1} = \begin{pmatrix} \hat{C}_{11} & \hat{C}_{12} & \hat{C}_{13} \\ \hat{C}'_{12} & \hat{C}_{22} & \hat{C}_{23} \\ \hat{C}'_{13} & \hat{C}'_{23} & \hat{C}_{33} \end{pmatrix} \quad (10)$$

For our set of single-tree equations there are  $p_1+p_2+p_3=8+8+10=26$  estimated parameters, plus the main diagonal partition of the variance-covariance matrix for each model ( $\hat{C}_1, \hat{C}_2, \hat{C}_3$ ). The remaining off diagonal partitions ( $\hat{C}_{12}, \hat{C}_{13}, \hat{C}_{23}$ ) contain the cross-equation covariances between each estimated parameter.

If  $X_1=X_2=X_3$  then SUR is identical to OLS, even when the cross-equation covariances are nonzero. Otherwise, SUR provides an asymptotically more precise estimator of  $\beta$  than OLS, and the comparative efficiency of  $\hat{\beta}$  increases with increasing cross equation correlation and with increasing dissimilarity among the regression matrices  $X_1, X_2$  and  $X_3$ .

The use of  $\ln(C)=Y_3$  in the models for  $\ln(BAI)$  and  $\ln(\Delta H)$ , eqns. (1) and (2) respectively, precludes the straightforward use of SUR as outlined. Whenever a response or endogenous variable from one model appears as a regressor variable in another model, both the OLS and SUR estimator of  $\beta$  will be biased and inconsistent (Judge et al. 1980). In this case two stage least squares or 2SLS (see Judge et al. 1980) can be used to provide consistent and asymptotically unbiased estimates of  $\beta$ .

An even more efficient estimator was invented by Zellner and Theil (1962): three-stage least square (3SLS). By combining Thiel's (1953) 2SLS procedure with Zellner's (1962) SUR procedure, the resulting 3SLS estimators of  $\beta$  are consistent and asymptotically more efficient than 2SLS. We denote the 3SLS estimates of  $\beta$  as  $\tilde{\beta}$  and its estimated variance as

$$\tilde{\Sigma} = c\tilde{\sigma}v(\tilde{\beta}) \quad (11)$$

It is convenient to partition  $\tilde{\Sigma}$  as

$$\tilde{\Sigma} = \begin{pmatrix} \tilde{C}_{11} & \tilde{C}_{12} & \tilde{C}_{13} \\ \tilde{C}'_{12} & \tilde{C}_{22} & \tilde{C}_{23} \\ \tilde{C}'_{13} & \tilde{C}'_{23} & \tilde{C}_{33} \end{pmatrix} \quad (12)$$

where  $\tilde{C}_{ij}$  is the  $p_j \times p_k$  estimated covariance matrix of  $\tilde{\beta}_j$  and  $\tilde{\beta}_k$ . For details concerning the 3SLS estimator in a forestry context see Murphy and Beltz (1981), Borders and Bailey (1986), and Gregoire (1987).

In an individual tree model, estimates of  $E[Y_1]$ ,  $E[Y_2]$ , and  $E[Y_3]$  are needed for each tree. Let  $x_1'$  denote the  $p_1 \times 1$  row vector of covariate values for a particular tree for which the estimated coefficients will be applied, such that

$$\tilde{y}_1 = x_1' \tilde{\beta}_1 \quad (13)$$

serve as an estimate of  $E[y_1, x_1']$ . Let  $\tilde{y}_2$  and  $\tilde{y}_3$  be similarly defined.

The distributional properties of the random errors  $\mathbf{e}$  in conjunction with the estimator of  $\beta_i$  determine the statistical properties of  $\tilde{\beta}_i$  and  $\tilde{y}_1$ . The scalar covariance between random variables  $\tilde{y}_1$  and  $\tilde{y}_2$  is  $x_1 \tilde{C}_{12} x_2'$ , where  $\tilde{C}_{12}$  is the  $p_1 \times p_2$  covariance matrix of  $\tilde{\beta}_1$  and  $\tilde{\beta}_2$ . Given an estimate of  $C_{12}$ , say,  $\tilde{C}_{12}$  as in (12), then the covariance between  $\tilde{y}_1$  and  $\tilde{y}_2$  can be estimated by  $x_1 \tilde{C}_{12} x_2'$ . Therefore, the correlation between  $\tilde{y}_1$  and  $\tilde{y}_2$  can be expressed as

$$\tilde{r}_{12} = \frac{x_1 \tilde{C}_{12} x_2'}{\sqrt{(x_1 \tilde{C}_{11} x_1')(x_2 \tilde{C}_{22} x_2')}} \quad (14)$$

The correlation  $\hat{r}_{13}$  between  $\hat{y}_1$  and  $\hat{y}_3$ , and the correlation  $\hat{r}_{23}$  between  $\hat{y}_2$  and  $\hat{y}_3$  can be estimated similarly.

## DATA

Data were obtained from the Austrian National Forest Inventory (Forstliche Bundesversuchsanstalt 1981), a systematic hidden permanent sample plot design over the whole of Austria, with a 5-yr remeasurement interval. In a given year a fifth of the plots are remeasured, ensuring a representative sample of all Austrian forests each year. The total inventory comprises 22,000 permanent plots. We restricted ourselves to the 4,135 forested plots containing remeasured Norway spruce (*Picea abies*), not crossed by roads, and in a single ownership.

Permanent sample plots were established from 1981 to 1985. Trees with a diameter at breast height (DBH, 1.3 m) larger than 10.4 cm were selected by angle count sampling using a BAF of 4 m<sup>2</sup>/ha. Trees with a DBH between 5 and 10.4 cm were measured within a circle of 2.6 m radius located at plot center; smaller trees were not recorded. At plot establishment, the following data were recorded for every sample tree: species, DBH to the nearest mm, and distance and azimuth from plot center. Total height and height to the crown base were measured to the nearest decimeter on every fifth tree. Plot descriptors were evaluated within a circle of 300 m<sup>2</sup>. Elevation is measured to the nearest 100 m, slope is measured to the nearest 10%, and aspect to the nearest 45°. Additional site descriptors were measured but not used in this comparison study.

Plots were remeasured from 1986 to 1990, 5 years after establishment. In the remeasurement, height and diameter were measured on the same trees, but not height to crown base. In summary, observations of 7,797 Norway spruce trees from 4,135 different permanent sample plots are available throughout Austria.

## RESULTS

### Coefficient estimates by OLS vs. 3SLS

Using the SYSLIN procedure in the Econometrics-Time Series module of

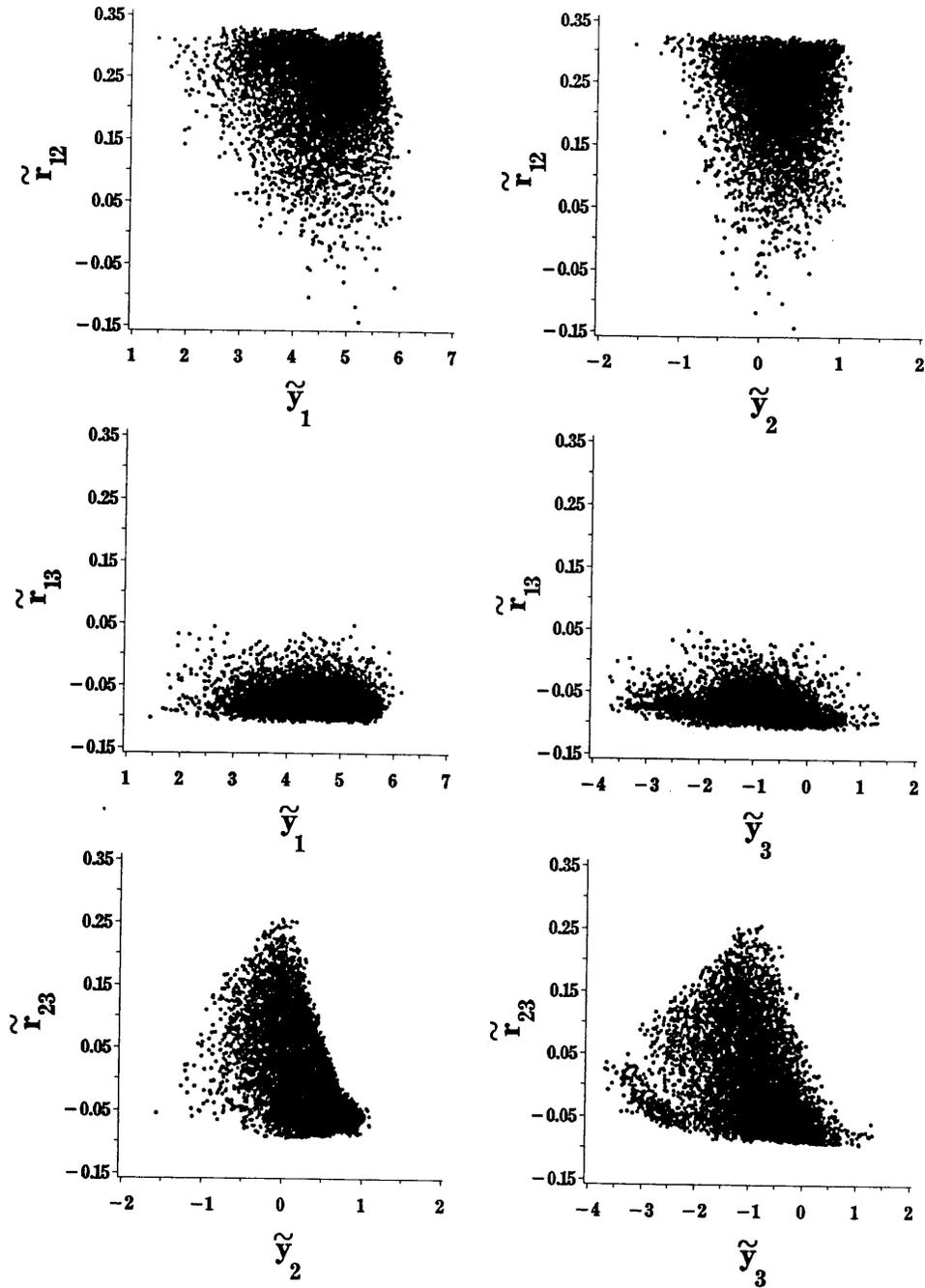
SAS (SAS Institute 1988), the parameters in equations (1) to (3) were first estimated independently by applying ordinary least squares (OLS) and then jointly by using 2SLS and 3SLS.

Attention was immediately focused on the  $\ln(C)$  and  $CCF$  terms in the height increment model (2). These two terms were both significant ( $\alpha=0.05$ ) with OLS. With 2SLS, the  $\ln(C)$  term remained significant but the  $CCF$  term became strongly non-significant. With 3SLS, both the  $\ln(C)$  and  $CCF$  terms were non-significant. Thus the height increment model (2) was reduced from 8 parameters to 6. The size of the matrix partitions in the variance-covariance matrix in (12) is reduced accordingly. Now  $\tilde{C}_{22}$  is a 6x6 variance-covariance matrix for  $\tilde{Y}_2$ , and the off-diagonal partitions  $\tilde{C}'_{12}$  and  $\tilde{C}'_{23}$  are 8x6 and 6x10 matrices containing the cross-equation covariances between the estimated parameters, respectively. Because there are now  $p_1+p_2+p_3=8+6+10=24$  estimated parameters in the three growth equations, there are 576 variance-covariance elements in (12), a reduction of 100 elements.

### Correlation between predictions

To investigate the strength of the interrelationships among our system of equations and evaluate their importance, the correlations between the predictions of  $\ln(\Delta BA)$ ,  $\ln(\Delta H)$  and  $\ln(C)$  (eqns. (1), (2), and (3)) are calculated for each observation according to equation (14).

Fig. 1 displays the cross-equation correlations  $\tilde{r}_{ij}$  between each pair of predictions vs. the respective predicted variables  $\tilde{Y}_i$ . In Fig. 1a and 1b,  $\tilde{r}_{12}$  is massed around 0.23, with a maximum correlation of 0.33 between the predictions for  $\ln(BAI)$  ( $\tilde{y}_1$ ) and  $\ln(\Delta H)$  ( $\tilde{y}_2$ ). This indicates that the first two equations in the system are fairly interdependent. The correlations  $\tilde{r}_{13}$  between predictions from the first ( $\tilde{y}_1$ ) and third equations ( $\tilde{y}_3$ ) are weakly correlated, with a mean of -0.07, and extrema at 0.05 and -0.11. The correlations  $\tilde{r}_{23}$  between predictions from the second ( $\tilde{y}_2$ ) and third equations ( $\tilde{y}_3$ ) have a mean of zero (0.01), but are more dispersed than those for  $\tilde{r}_{12}$ , ranging from 0.26 to -0.10 with a standard deviation of 0.08. These correlations between the basal area and height increment models and between the height increment and crown models confirm our assumption of cross-equation correlations.



**Fig. 1:** The cross equation correlation  $\tilde{r}_{ij}$  between each pair of  $\tilde{y}_i$  predictions;  $\tilde{y}_1$  indicates the predicted  $\ln(\Delta BA)$ ,  $\tilde{y}_2$  the predicted  $\ln(\Delta H)$ , and  $\tilde{y}_3$  the predicted  $\ln((1/CR)-1)$ .

## DISCUSSION

Because individual tree forest growth models are based on multivariate attributes observed on the same individuals (e.g. basal area increment, height increment, crown ratio), the resulting set of growth equations can be considered a simultaneous system. Therefore, joint regression techniques should be considered for simultaneously determining the parameter estimates. If endogenous variables do not appear on the right hand side (RHS), the seemingly unrelated regression (Zellner 1962) will still improve the efficiency of the parameter estimates. If endogenous variables are used as predictor variables then multi-stage estimation techniques (2SLS or 3SLS) are necessary to obtain parameter estimates that are consistent; such estimates will also be asymptotically efficient. Ignoring the simultaneous nature of the system by separately applying OLS to each model can result in estimates that are biased and inconsistent.

Using 3SLS with our system of equations allowed the detection and deletion of two non-significant terms ( $CCF$  and  $\ln(C)$  in eq. (2)) that OLS had determined as significant. This led to a simplification of the simultaneous structure of our system because  $\ln(C)$  was originally considered as an endogenous variable on the RHS.

One of the advantages of joint regression techniques is that both the correlation among the predictor variables within a certain equation and across all equations are available. Fig. 1 indicates that cross-equation may change their degree of dependency in the system and that these correlations can be used in a simulator to better predict stand and tree dynamics. Predictions between the first two growth equations (basal area and height increment) are fairly well correlated, while correlations between predictions from the first and third (height increment and crown ratio) are rather weak.

## ACKNOWLEDGEMENTS

This research was conducted when Hubert Hasenauer was a Visiting Scientist at Virginia Tech Department of Forestry in Blacksburg, and at the Intermountain Research Station's Forestry Sciences Laboratory in Moscow, Idaho. Hasenauer was working on a Schrödinger research grant from the Austrian Science Foundation. We are grateful to Karl Schieler and Klemens Schadauer of the Federal Forest Research Center in Vienna for making the Forest Inventory data available.

## LITERATURE

- Aitken, A.C. 1934-35. On the least-squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh*. 55: 42-48.
- Borders, B.E., and R.L. Bailey. 1986. A compatible system of growth and yield equations for slash pine fitted with restricted tree stage least squares. *For. Sci.* 32: 185-201.
- Burkhart, H.E., Farrar, K.D., Amateis, R.A., and Daniels, R.F., 1987. Simulation of individual tree growth and stand development in loblolly pine plantations on cutover, site-prepared areas. *Coll. For. and Wildlife Resources*. Virg. Tech. Inst., Blacksburg, Va. Publication FWS-1-87: 47 pp.
- Dixon, R.K., Meldahl, R.S., Ruark, G.A., and Warren, W.G. 1990. Process modeling of forest growth responses to environmental stress. Timber Press, Portland, OR. 441 pp.
- Forstliche Bundesversuchsanstalt, 1981. Instruktionen für die Feldarbeit der Österreichischen Forstinventur 1981-1985. Forstliche Bundesversuchsanstalt, Wien, 172 pp.
- Gregoire, T.G. 1987. Generalized error structure for forestry yield models. *For. Sci.* 33: 423-444.
- Haavelmo, T. 1943. The statistical implications of a system of simultaneous equations. *Econometrica* 11: 1-12.
- Hasenauer, H. 1994. Ein Einzelbaumwachstumssimulator für ungleichaltrige Fichten-Kiefern- und Buchen-Fichtenmischbestände. *Forstl. Schriftenreihe, Univ. f. Bodenkultur, Wien. Österr. Ges. f. Waldökosystemforschung und experimentelle Baumforschung an der Univ. f. Bodenkultur. Band 8: 152 pp.*
- Hasenauer, H., and R.A. Monserud. 1996a. Biased statistics from height increment predictions. (submitted to *Ecol. Modeling*).
- Hasenauer, H., and R.A. Monserud. 1996b. A crown ratio model for Austrian forests. *Forest Ecology and Management* (in press).
- Judge, G.G., W.E. Griffiths, R.C. Hill, and T.C. Lee. 1980. The theory and practice of econometrics. John Wiley & Sons, Inc., New York. 810 pp.
- Krajicek, J.E., K.A. Brinkman, and S.F. Gingrich. 1961. Crown competition: A measure of density. *For. Sci.* 7: 35-42.
- Monserud, R.A. 1975. Methodology for simulating Wisconsin northern hardwood stand dynamics. Ph.D. Thesis, Univ. of Wisconsin, Madison. 156 pp.
- Monserud, R.A., and H. Sterba. 1996. A basal area increment model for individual trees growing in even- and uneven-aged forest stands in Austria. *For. Ecol. Manage.* (in press).
- Murphy, P.A., and R.C. Beltz. 1981. Growth and yield of shortleaf pine in the West Gulf region. South. Forest Exp. Station, New Orleans, LA. USDA For.Serv.Res.Paper SO-169: 15 pp.
- SAS Institute. 1988. SAS/ETS user's guide, version 6. Cary, NC. 560 pp.
- Theil, H. 1953. Repeated least-squares applied to complete equation systems. The Hague: The central planing bureau. The Netherlands.
- Wykoff, W.R., Crookston, N.L. and Stage, A.R., 1982. User's Guide to the Stand Prognosis Model. USDA For. Serv. GTR INT-133, 112 pp.
- Zellner, A. 1962. An Efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Stat. Association.* 57: 348-368.
- Zellner, A., and Theil, H. 1962. Three-stage least squares: simultaneous estimation of simultaneous equations. *Econometrica* 30: 54-78.

# Genetics, geographics, and prairie dogs: error and accuracy in a validated spatially-explicit dispersal model.

Gillian Bowser<sup>1</sup>

---

**Abstract.**— Genetic theory predicts that populations will differentiate based on the geographic distance between them unless some environmental feature creates a barrier to dispersal. However, those barriers are often inferred from landscape features, and rarely directly tested for their ability to predict genetic differentiation as expressed by genetic distances. There is a need to develop spatially explicit models validated by population characteristics, such as genetics, to predict population sub-division—an important step in the conservation of a species. Prairie dogs (*Cynomys ludovicianus*), are colonial fossorial rodents that live in discrete populations which have distinctive signatures in remotely-sensed images, providing a mechanism to combine spatial and genetic databases into a single model. However, this combination of spatial and genetic data complicates spatial error as spatial genetic parameters, measured from protein differentiation, have associated error around distance algorithms; while spatial databases, such as elevation, hydrology, and transportation, have error associated with placement. The implications of testing spatial models on small organisms at the landscape level are discussed

---

## INTRODUCTION

Ecological modeling of animal populations with remotely-sensed data is a relatively new field for conservation biologists (Johnson 1990). How animals move among populations and how different environmental features affect population exchange and stability are fundamental questions for the conservation of patchy fragmented habitats (Johnson and Naiman 1987, Jensen et al. 1990). Given the current availability of remotely-sensed databases along with the statistical tools to analyze those databases, there is also an increasing need to address the accuracy and resolution of these landscape databases when used at the scale of an individual animal (Costanza and Maxwell 1994).

Modeling population establishment and movement on a landscape level requires using some remotely-sensed data (such as vegetation, elevation, or the like) to imply where the animal populations may occur (Scott et al. 1993). However, the digital databases expressing the environmental parameters contain sources of error associated

---

<sup>1</sup>Natural Resource Specialist, Rocky Mountain System Support Office, National Park Service, Denver, Colorado

with the scale, resolution and projection of the source map that can have important implications when applied to ecological databases (Meenemeyer and Box 1987, Lee et al. 1992, Kemp 1993, Bolstad and Stowe 1994, Hodgson 1995). In addition, the unit of scale of the landscape will change with the different ecological processes being examined and will vary with different organisms as to the correct scale for expressing spatial movements (Hunsacker et al. 1993 Cale and Hobbs 1994). At a certain point, the error associated with the scale of the underlying database can overwhelm any potential to model populations at finer resolutions (Brown and Bara 1994, Berry 1995). This point-of-no-return can be expressed as an error term or through error management to evaluate model performance (Hunter and Goodchild 1995).

Here I discuss the implications of scale and resolution on constructing a spatially explicit model of animal dispersal, and how the validation of this model is limited by discrepancies between the ecological scale (how far an individual animal can move on a landscape) and the database scale (what are the limitations of resolution and projection on the database accuracy for modeling the movements of small mammals).

## **BACKGROUND**

Dispersal among populations is essential to maintain genetic homogeneity within a population thus preserving population health and longevity (Green 1994). As populations become isolated through environmental or behavioral barriers, genetic divergence occurs leading eventually towards speciation. The progress towards speciation can be estimated on a shorter time scale by calculating the dispersal rates per generation and creating a genetic distance (based on exchange as mirrored by genetic differentiation-- Nei 1972) that can validate dispersal distances (a rate of exchange based on physical separation).

To examine dispersal distances across realistic landscapes as a measure of population conservation, each element of a landscape (elevation, vegetation, transportation, hydrology) can be viewed as an independent treatments (single model runs) to be tested against the predicted dispersal distances based on genetic differentiation (as a measure of population persistence. The effects of environment elements (such as roads or developments) on dispersing animals can be directly by calculating population parameters (genetic differentiation) against an estimate of the exchange rate (a dispersal distance) that creates a spatially-explicit dispersal model applicable for population conservation efforts.

### **Model Parameters**

Three aspects of ecological model are directly affected by the scale and resolution of the spatial databases themselves: the size of the area covered by the model; the size of the species the model is designed to emulate; and the variability in the landscape.

#### **How big is the study area?**

The first step in model construction is to define the size of the patch addressed by the model. If the patch is a preserve or politically defined boundary, there may be constraints defined by the boundary of that patch which should be included in any

modeling efforts. In addition, the size of the total area of consideration can determine the resolution of the databases feasible given hardware and software limitations. Highly mobile insects, for example, would require extremely high resolution data over large areas that is beyond the storage capacity of many computers.

#### **How large is the organism?**

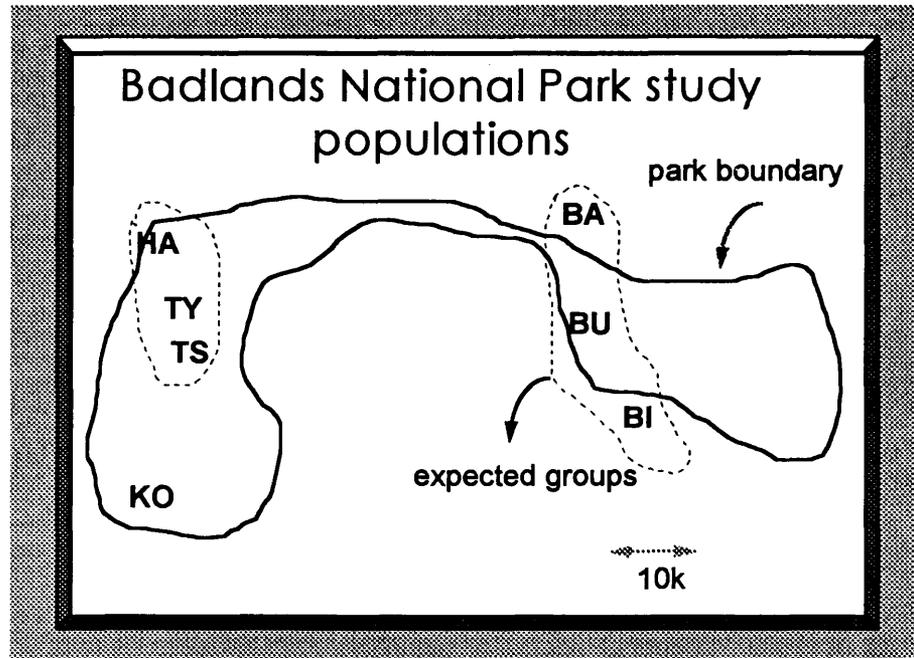
The size of the species will dictate the required precision of any underlying map layers (Brown and Bara 1994). Smaller species may be more sensitive to the resolution of a spatial data than larger species and thus a model based on a coarser resolution data could miss- or over-represent landscape features and their impact on animal movement. The physical size of the species will also dictate how rapidly dispersal declines with geographic distance, in which case modeling population placement may be more valid than individual movement (even though this is fundamentally a different question) and the resolution of digital layers may be unrealistic for that species (Cale and Hobbs 1994).

#### **How variable is the study area?**

Habitat variability, especially as it affects small animals, can be under-represented by databases with coarse resolution. For example, small erosive features, long narrow gullies or streambeds, may not be visible to a database with a minimum pixel size of 30 meters. If those gullies are barriers to the movement of small animals, the database may not be able to detect those barriers and could mis-represent animal movement.

### **STUDY SITE**

Black-tailed prairie dogs (*Cynomys ludovicianus*) are colonial fossorial rodents of the Great Plains of North America. The prairie dog is considered a keystone species for the North American plains because they form large colonies and actively modify the environment within their colonies, creating unique ecosystems. The colonies with their distinct vegetation are distinct in aerial photography (Schenbeck and Myre 1990) and thus are ideal to demonstrate spatially-explicit dispersal models. Badlands National Park, South Dakota has a number of large prairie dog colonies scattered throughout the park. As the park has two distinct lobes with colonies on either end, predicted dispersal patterns among colonies are colonies relatively close on one side of the park would be genetically more similar than those on opposite (figure 1). These predictions (euclidean distances describe population dispersal patterns) will serve to validate model predictions based on the dispersal and genetic distances.



**Figure 1.** Cartoon of the distribution of major prairie dog populations in Badlands National Park, South Dakota. Population abbreviations are: Baysinger (BA), Burns Basin (BU), Bigfoot (BI), Haybutte (HA), Kocher Flats (KO), Tyree Basin (TY), Tyree South (TS).

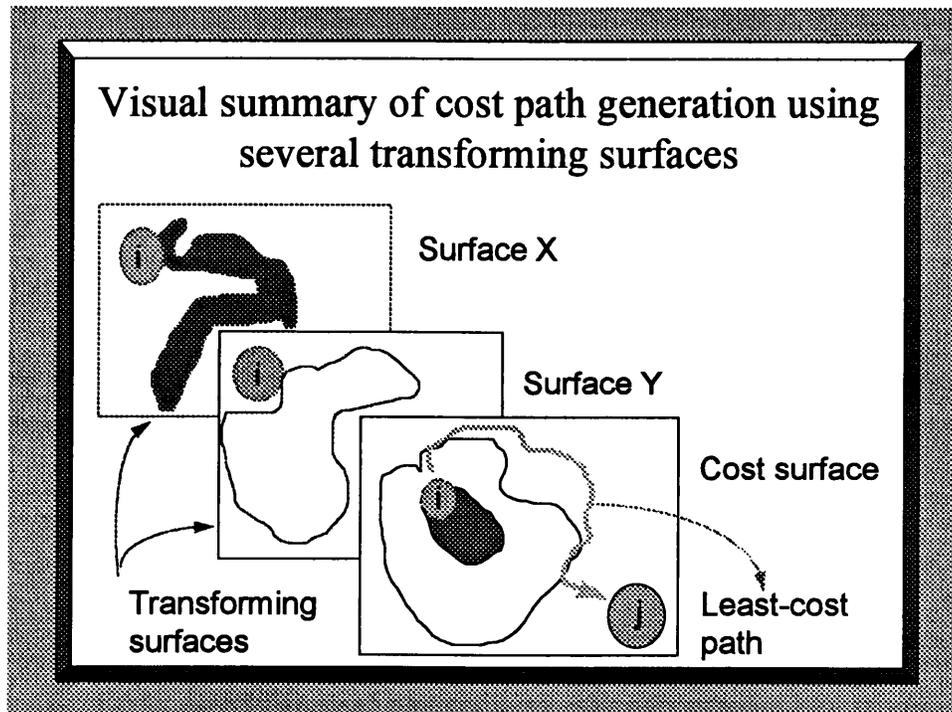
## MODEL CONSTRUCTION

Environmental layers such as Digital Elevation Model (DEM-1:24,000), Digital Line Graph (DLG--1:250,000), aerial photography (black and white and color infra-red: 1:24000), Soil Conservation Service (SCS DLG--1:250,000), vegetation (National Park Service (NPS) database--1:24000) and landuse boundaries (NPS database-1:250,000) were entered into a GRASS4.1 GIS database on a Sun Sparkstation. All data were rectified with GRASS program modules for geo-referencing (Clark66) and spheroid corrections (NAD27). Additional corrections for resolution were done using C- and Bourne-shell scripts and GRASS resampling. All model runs were generated by these scripts and were run in the UNIX environment, calling the GRASS modules as needed.

A model run consisted of an individual GRASS surface (such as roads) redefined based on user-assigned weights meant to symbolize the ease or difficulty for prairie dogs to cross a particular environmental feature. Those weights are then expressed as a cost surface (the cumulative weights for moving from a particular point to any other point on the surface) which is then used to generate a dispersal path (figure 2). Dispersal paths are calculated using algorithms from hydrology as water seeks the lowest path down a drainage. The route water would follow then represents the least-cost path that an animal would follow over an ecologically-defined surface (figure 3).

Dispersal paths were validated by using the genetic differentiation of proteins (as determined by the horizontal gel-electrophoresis of polymorphic enzymes) to create a genetic distance analogous to the physical separation (see Bowser 1996 for methods). Genetic distance is a metric measure of the genetic differentiation of proteins among

population samples, and can be used as an independent measure to test for linear associations with the paths generated by the dispersal models (Nei 1972, Slatkin 1985). Genetic distances were calculated using BIOSYS-1 (Swofford and Selander 1989) and Weirs (1990) jack-knifing procedure.



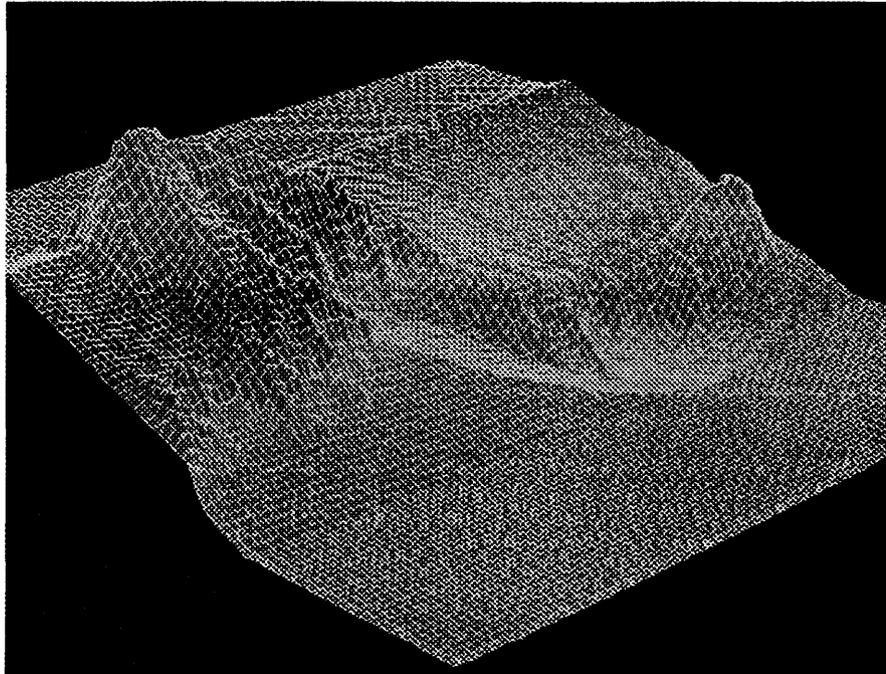
**Figure 2.** Conceptual diagram of the cost path generation using surfaces consisting of single environmental feature to create composite surfaces. Populations are represented by 'i' and 'j' and paths are calculated from the centroids of i and j. Different polygons indicate ecological features on the transforming surface that pose a barrier to dispersal.

All statistical tests were done using Splus3.3 in the UNIX environment after corrections for normality and checks for autocorrelation. Each environmental surface was seen as an independent model consisting of a single parameter such as roads, streams, etc. Composite layers were created within GIS from combinations of single layers which were then rescaled to account for changes in resolution and tested as a single independent surface

## RESULTS

Dispersal paths calculated for prairie dog populations in BADL were significantly different than the Euclidean distances among the populations (table 1). All of the environmental surfaces created longer paths than the Euclidean distances with the exception of vegetation which was not significantly different. Loglinear regression on splined distances showed that there were weak predictive relationships ( $p=0.05$ ,  $r^2=.14$ ) for single models between the genetic distances and the dispersal distances (figure 4). The best fitted model using step-wise regression was roads, streams, and park boundary

( $p=0.05$ ,  $r^2=0.33$ ). Composite models were generally better at predicting the genetic distances than single models using Principal Component Analysis; composite models contributed the most to the first PCA axis which explained 86% of the variance in genetic distances.



**Figure 3.** Cost surface for moving from one location (the pit) to any other point on the surface. Cost refers to the cumulative sum of pixel weights from a designated starting point to any other point on the surface.

**Table 1.** ANOVA on single surface models as independent treatments against Euclidean distance among populations of prairie dogs in Badlands National Park.

Model	df	f	p	sign
roads	1	64.55	0.00	**
streams	1	75332.6	0.00	**
slope	1	159.97	0.00	**
elevation	1	286.23	0.00	**
vegetation	1	1.20	0.28	NS

## DISCUSSION

The validated dispersal models demonstrated here showed an ability to predict another population parameter, genetic distance, which is an important measure of the longevity, exchange, and isolation of populations. As an exploratory model, the dispersal model was able to demonstrate which environmental layer appear to best explain the genetic patterns observed, as well as provide some insights as to which environmental features may have little effect. Despite the presence of the badlands wall and other erosion features within BADL, elevation was not a strong predictor of genetic distance.

In contrast, streams and roads together as a composite model was best able to explain the genetic patterns over any single environmental feature.

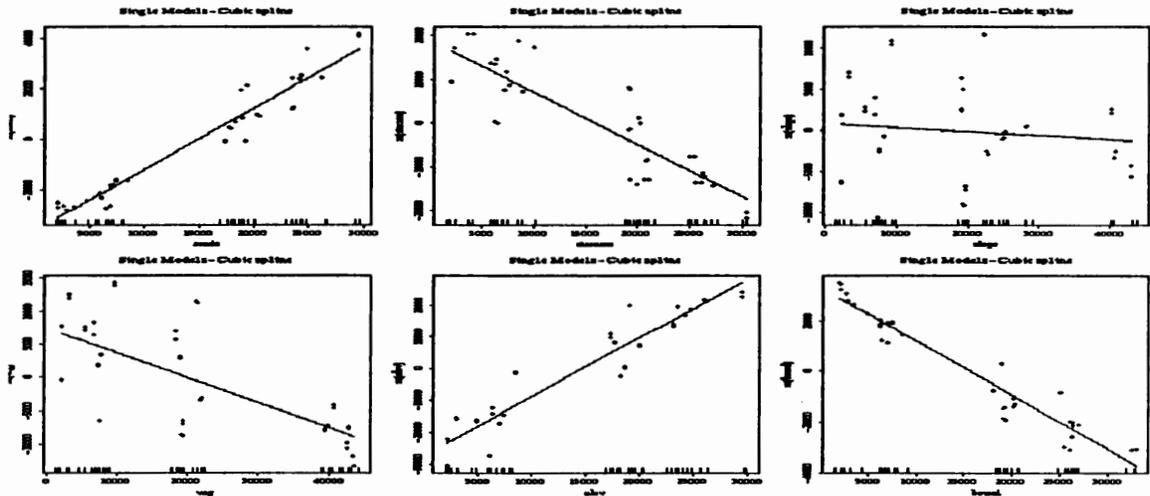


Figure 3. Regression lines for each single surface tested using genetic distances as the dependent variable. Graphs labels from the left: roads, streams, slope, vegetation, elevation, and park boundary.

### The Influence of Error

There are two sources of error discussed here: error associated with the placement of populations; and error associated with treatment affects during the model validation.

#### Population placement

The location of populations on the landscape for a dispersal model is critical as the seed point for constructing paths on the cost surface. If a centroid position for a population polygon is used as the starting point, the error associated with that centroid only becomes important when referenced to ground measurements. Each centroid has a sphere of error round it that represents the corrections due to population placement (digitizing and interpretation error), RMS associated with the photograph rectification, along with any error associated with

#### Treatment error

Treatment error builds on existing population placement error within the framework of a statistical test. Location error can be analogous to the variance around each point within a treatment. Testing the variance among treatments (assuming each treatment is a different spatial layer), changes the total error incorporated within the statistical model would be a combination of the variance among and within treatments--the first due to population placement and the second due to the underlying precision and resolution of the original environmental database. The combined errors reduce the power of a variance test to explain the "background noise" resulting in low multiple  $r$  squared values.

### CONCLUSIONS

Error, scale, and resolution can limit the ability of GIS to accurately model animal

movements across complex landscapes. When a model is validated by a population measure (such as genetic distances) the error incorporated into spatial models from digital databases and the assumptions of accuracy underlying those databases are elucidated. In the dispersal model, the inability to explain more than 33% of the model variance in the regression models yet still have a demonstrable relationship (and reproduce patterns expected by the population parameters) suggests that the error in placement may distort model results. The take-home message is that although the dispersal model did a modest job in re-creating the genetic patterns and sorting through which environmental features best explained those patterns, the accuracy, precision, and resolution of the underlying databases must be discussed. In this model, the minimum resolution (30 meters) was greater than most of the animal's home territories (and some colonies). Realistically, the model requires the animal to move in jumps of at least one meter up to 30 meters depending on the surface--a difficult task for an animal of 2-10 pounds. The fact that the model still distinguished among surfaces and replicated genetic distance predictions is intriguing and will hopefully generated further exploration as to the effects of the accuracy and error of spatial databases on ecological modeling.

### ACKNOWLEDGEMENTS

Bette Loiselle and John Blake, my dissertation advisors from the University of Missouri-St. Louis, provided strong support in the development of this project as a part of my dissertation. Chip Harvey, Chris Theriault, and Michelle Gudorf of the National Park Service provided excellent discussions on spatial error and ecological modeling. I appreciate the efforts of Peter Strong, David Duran, and Bruce Powell of the National Biological Service (NBS) to keep my computer running. Ralph Root and Susan Stitt (NBS) provided excellent analytical support for the spatial analysis.

### LITERATURE CITED

- Bolstad, P. V. and T. Stowe 1994. An evaluation of DEM accuracy: elevation, slope and aspect. *Photogrammetric Engineering and Remote Sensing* 60(11): 1327-1332.
- Costanza, R. and T. Maxwell 1994. Resolution and predictability: an approach to the scaling problem. *Landscape Ecology* 9(1): 47-57.
- Bowser, G. 1996. Integrating ecological tools with remote sensing: modeling animal dispersal on complex landscapes. Conference proceeding Third Annual Conference on Environmental Modeling and GIS, Santa Fe, New Mexico, CDROM.
- Brown, D. G. and T. J. Bara 1994. Recognition and reduction of systematic error in elevation and derivative surfaces from 7 1/2 minute DEMs. *Photogrammetric Engineering and Remote Sensing* 60(2): 189-194
- Goodchild, M. F. 1993. The state of GIS for environmental problem solving *in Environmental Modeling with GIS*. M.F. Goodchild, B. O. Parks and L. T. Steyaert editors. Oxford University Press. 485 pages
- Green, D. G. 1994. Connectivity and complexity in landscapes and ecosystems. *Pacific Conservation Biology* 1:194-200.

- Hodgson, M. E. 1995 What cell size does the computed slope/aspect represent? *Photogrammetric Engineering and Remote Sensing* 61(5): 513-517.
- Hunter, G. J. and M. F. Goodchild 1995. Dealing with error in spatial databases: a simple case study. *Photogrammetric Engineering and Remote Sensing* 61(5): 529-537.
- Jensen, J.R., S, Narumalani, O. Weatherbee and K. M. 1992. Predictive modeling of cattails and waterlily distribution in a South Carolina reservoir using GIS. *Photogrammetric Engineering and Remote Sensing* 58: 1561-1568.
- Johnson, C. A. and R. J. Naiman 1990. The use of geographical information systems to analyze long-term landscape alteration by beaver. *Landscape Ecology* 4: 5-19.
- Johnson, L. B. 1990. Analyzing spatial and temporal phenomena using geographic information systems: A review of ecological applications. *Landscape Ecology* 4: 31-43
- Hunsaker, C. T., R. A. Nisbet, D. C. L. Lam, J. A. Browder, W. L. Baker, M. G. Turner, and D. B. Botkin. 1993. Spatial model of ecological systems and processes: the role of GIS. *in Environmental Modeling with GIS*. M.F. Goodchild, B. O. Parks and L. T. Steyaert editors. Oxford University Press. 485 pages
- Kemp, K. K. 1993. Spatial databases: sources and issues. *in Environmental Modeling with GIS*. M.F. Goodchild, B. O. Parks and L. T. Steyaert editors. Oxford University Press. 485 pages.
- Lee, L., P. K. Snyder, and P. F. Fisher 1992. Modeling the effect of data errors on feature extraction from digital elevation models. *Photogrammetric Engineering and Remote Sensing* 58(10): 1461-1467.
- Schenbeck, G. L. and R. J. Myhre 1986. Aerial photography for assessment of black-tailed prairie dog management on the Buffalo Gap National Grassland, South Dakota. USDA Report no 86-7 3400.
- Scott, J. M., F. Davis, B. Csuti, R. Noss, B. Butterfield, C. Groves, H. Anderson, S. Caicco, F. D'Erchia, T. C. Edwards jr, J. Ulliman, and R. G. Wright 1993. GAP analysis: a geographic approach to protection of biological diversity. *Wildlife Monographs* 123.
- Swofford, D. L. and R. B. Selander 1989. Biosys-1: A computer program fo the analysis of allelic variation in population genetics and biochemical systematics. Release 1.7 Illinois Natural History Association, Illinois.
- Wickham, J. D. and D. J. Norton 1994. Mapping and analyzing landscape patterns. *Landscape ecology* 9(1): 7-23.
- Wier, B. 1990. Genetic Data Analysis. Sinauer Associates, New York

## **BIOGRAPHICAL SKETCH**

Gillian Bowser is a natural resource specialist with the National Park Service in Denver, Colorado. She holds an MS from the University of Vermont in Zoology and is currently a PhD candidate with the University of Missouri-St. Louis. Gillian provides technical assistance and develops proposals in the sciences for national parks in the rocky mountain area.

# Effect of Uncertainty in Mapped Biodiversity Data on Optimal Conservation Decisions

Michael J. Conroy<sup>1</sup> and Jennifer E. Crocker<sup>2</sup>

---

**Abstract**--Spatially referenced biological diversity data are increasingly being used to assist with conservation decisions. Examples include the identification, selection, and design of ecological reserves; the types, amount, and distribution of forest cutting; and the placement of corridors to connect reserves. These have in common: (1) an implicit long-term objective; (2) dynamic systems; (3) alternative possible decisions; (4) alternative uncertain outcomes; (5) uncertainty about both the current system state (sampling errors), and in predictions about future system states ("model uncertainty"). Decision-theoretic methods and predictive models can be used to find decisions that result in a long term optimum (e.g., a balance between species conservation and timber revenue), subject to updating through current surveys, research, and other information. We illustrate the approach for a hypothetical reserve design problem involving four species with different habitat affinities in a landscape with two habitat types, with both demographic and statistical uncertainty. In particular, we examine how uncertainty in predictive models (e.g., models relating animal distribution and abundance to habitat features) can result in suboptimal decisions, and how research and monitoring can be used adaptively to reduce uncertainty and improve decision making.

---

## INTRODUCTION

Increasingly, managers are faced with making decisions about how to manage systems in which there are multiple, competing goals or objectives, and the system under management is both complex and incompletely understood. The failure of public decision makers to make a "good" decision can have severe consequences, ranging from loss of agency credibility and good will, to unnecessary losses of revenue or opportunity, to adverse impacts on the resource, including the loss of species. Concern over loss of biological diversity has led to efforts in the

---

<sup>1</sup> Assistant Leader, Georgia Cooperative Fish and Wildlife Research Unit, Athens, GA.

<sup>2</sup> Graduate research assistant, Institute of Ecology, University of Georgia, Athens, GA

cartographic analyses of species distributions and landscape characteristics (e.g., Burley 1988; Scott et al. 1993) In these efforts, animal species richness is either observed directly or inferred from vegetation (habitat) maps and models relating abundance or presence/absence to vegetation cover attributes and physical features (Scott et al. 1993). The mapped data are then used land-use decisions directed at the conservation of biological diversity, such as reserve design (Scott et al. 1993).

Biodiversity mapping programs such as Gap Analysis frequently do not formally incorporate uncertainty about 1) mapping errors in habitat and animal species presence distributions, and 2) species-habitat associations. They implicitly assume that predictions can be made about the future state of biodiversity, given alterations in current habitat conditions (e.g., through management). Here we consider how statistical unreliability in spatially referenced data may affect the utility of mapped species richness patterns for conservation decisions such as reserve selection and design. We first briefly consider alternative approaches to obtaining optimal management decisions under the conditions in which the system is 1) stochastic, 2) dynamic, 3) imperfectly understood, and 4) incompletely observable. We illustrate the general problem by means of an artificial landscape containing 2 habitat types and 4 species (Conroy and Noon in press, Crocker and Conroy 1995).

Before proceeding, we note that we are using one specific method (stochastic dynamic programming) for determining optimal policies. There are a variety of alternative analytical approaches to this problem, capably reviewed by Williams (1989). These include non-dynamic methods such as linear programming, simulation methods, and combinations of simulation and optimization. We recognize that no single approach is likely to perform adequately under all circumstances, but we strongly feel that any approach should adequately reflect both the dynamic nature of ecological systems and decision making, and the role of uncertainty in both observations and model structure. We also agree with Walters (1986) and Nichols et al. (1995) that a formally adaptive approach to natural resource management will in the long run prove most efficacious.

## **METHODS**

### **Elements of the Problem**

We consider a hypothetical 100 hectare landscape with three cover types, which we have termed forest, open, and developed. Initially, the landscape is 75% forested, 25% open, and 0% developed. This landscape is initially occupied by four hypothetical species, two forest habitat specialists and two open habitat specialists. The management objective expresses a tradeoff between the conservation goal of maintaining the species in the landscape and the economic goal of allowing development. At five year intervals, the manager must decide what proportion of each habitat type to conserve, and a proportion of all unconserved land is

permanently lost to development. The manager has a range of hypotheses regarding population response to habitat management. The population is monitored during the management period, both to guide management decisions and to reduce uncertainties about population dynamics. In the development below we freely borrow from Nichols et. al. (1995), whom we thank for their lucid description of the problem.

### Management options

In this simplistic, non-spatial problem, the manager must decide how much of the available wildlife forest and open habitat to conserve for the next five year period. The action chosen in period  $t$  is denoted  $a_t$ . A policy  $A$  specifies a sequence of management actions  $\{a_0, a_1, \dots, a_n\}$  as a function of the observed population and habitat conditions.

### Models

The models specify the dynamics of the state variables through time. The set of population models specifies a range of alternative hypotheses about the population responses to management actions. The population dynamics and environmental dynamics are combined in a single state dynamics equation,

$$x_{t+1} = x_t + F_i(x_t, a_t, z_t)$$

where  $x_t$  is a vector of state variables including both population status and habitat status,  $z_t$  is a white noise process, and  $F_i(x_t, a_t, z_t)$  is the net change in the state variables under model  $i$ .

We considered three models describing population response to management actions. These models differ in the degree to which population dynamics is related to the amount, composition, and spatial arrangement of habitat. Model 1 was a simple non-dynamic habitat relation model,  $N=dH$ , where  $N$ =species abundance,  $d$ =a constant density,  $H$ =the amount of habitat (forest or open, depending on the species).

Model 2 was an equilibrium source-sink model (Pulliam 1988). For each species,

$$N = Hd \left( 1 + \frac{\lambda_1 - 1}{1 - \lambda_2} \right)$$

where  $N$ =species abundance,  $H$ =amount of source habitat,  $d$ =maximum density in the source habitat,  $\lambda_1$ = population growth rate in the source habitat, and  $\lambda_2$  = population growth rate in the sink habitat.

Model 3 was a dynamic source-sink model (Pulliam 1988). For each species, dynamics were described by:

$$N(1)_{t+1} = \begin{cases} \lambda(1)N(1)_t & \text{if } \lambda(1)N(1)_t < dH \\ dH & \text{if } \lambda(1)N(1)_t \geq dH \end{cases}$$

$$N(2)_{t+1} = \begin{cases} 0 & \text{if } \lambda(1)N(1)_t < dH \\ \lambda(2)N(2)_t + \lambda(1)N(1)_t - dH & \text{if } \lambda(1)N(1)_t \geq dH \end{cases}$$

where  $N(1)$ =population in source,  $\lambda(1)$ =population growth rate in source,  $d$ = maximum density in source,  $H$ = amount of source habitat,  $N(2)$ =population in sink,  $\lambda(2)$  = population growth rate in sink. Values of the species parameters are given in Table 1.

**Table 1.--Model parameters for population dynamics.**

	Habitat affinity	$\lambda_1$	$\lambda_2$	$d$
Species 1	Forest	1.3	1.01	1
Species 2	Forest	0.85	0.01	0.7
Species 3	Open	1.3	0.85	9
Species 4	Open	1.01	0.01	6

More complicated, spatially-explicit models can and will be constructed, but for the present Models 2 and 3 allow dispersal from favorable to unfavorable habitats, and thus are 'spatially informed', whereas for Model 1, in which movement between habitats does not occur, it is only the amount of favorable habitat that determines abundance for each species.

Landscape dynamics followed a simple state transition model. Unconserved land was permanently lost to development at a rate of 0.049 every 5 year period. Of the remaining habitat, forest converted to open at a 5 year rate of 0.26, and open to forest at a 5 year rate of 0.95.

### Components of uncertainty

There are four components of uncertainty which are important in managing biological systems. Environmental stochasticity is uncontrollable variation in the factors influencing population dynamics. We modelled extinction as a stochastic process, with probability distributions conditioned on population size. Structural uncertainty reflects imperfect knowledge about the relationship between management actions and population status. We explored the effect of structural uncertainty by using a weighted average of the three model predictions in the optimization procedure, and varying the weights assigned to each model. Partial observability refers to the imprecision in a monitoring program. Such imprecision leads to uncertainty about the appropriate action to take, and also can affect the ability to discriminate among alternative models. We plan to model the effect of partial

observability by developing an optimal policy based on complete observability, and then using that policy in a simulation model with random variation around the population state variables. Partial controllability refers to the inability to implement management decisions without error. We have not considered partial controllability in this exercise, but plan to do so in future work.

### Monitoring

Since decisions are partially based on population status, a monitoring program is needed. The monitoring program can also serve to discriminate among the competing models. The data recorded at time  $t$  are denoted  $y_t$ , and are related to the population according to

$$y_t = g(x_t) + \epsilon_t$$

with the random variable  $\epsilon$  independent of  $z$ . The set of all data collected up to time  $t$  is denoted  $Y_t$ .

### Model weights

The likelihood at time  $t$  that model  $i$  is the best model relating population to habitat is denoted  $p_i(t)$ . The likelihoods are expected to change through time as data are accumulated and compared with model predictions.

### Objective function

Our objective function  $V$  is a weighted sum of a species conservation objective and a resource use objective. In practice, the weights would be determined using multi-attribute utility theory (Keeney and Raiffa 1976). Our objective is to maximize

$$V(A|Y_0, p_0) = \sum_i p_0(i) E(W_1 S(A) + W_2 (1 - C))$$

where  $V(A|Y_0, p_0)$  is the expected value of the objective for policy  $A$  given the current state of knowledge,  $W_1 = 0.1875$  and  $W_2 = 0.25$  are weights for the two objectives,  $S(A)$  is the number of species at the end of 100 years, and  $C$  is the proportion of the available habitat conserved over all time steps. The weights were chosen so that the objective function falls between 0 and 1.

### Solution method

All optimizations and simulations were done using Stochastic Dynamic Programming (Lubow 1993), which uses backward iteration to find optimal policies for discrete Markov processes.

## Analysis

Our analysis focused on the effect of structural uncertainty on the optimal policy and the expected value of the objective function. Seven combinations of weights  $p_i$  were used (Table 2), reflecting a range of likelihoods ascribed to the three models. The first three cases reflected structural certainty. In the next three cases there is greater confidence in one model than the other two, and in the third there is no prior knowledge about the appropriate model.

In the extreme case in which the data  $y_t$  indicate with certainty which model is correct, we obtain perfect information (Lindley 1985) about the system structure. Given prior weights  $p_i$  given to the models, the expected value of the objective function with perfect information is:

$$\sum_{i=1}^3 \max_A V(A)p_i$$

The difference between this value and the maximum value of the objective function calculated using the weighted (by the  $p_i$ ) average of the three models is the expected value of perfect information (EVPI) (Lindley 1985). The EVPI was calculated for the four combinations of weights which reflected structural uncertainty.

## RESULTS

**Table 2.--The effect of varying model weights on optimal decisions and objective  $V$**

$p_1$	$p_2$	$p_3$	$V$	EVPI	forest area conserved	open area conserved
1	0	0	0.46	---	75	25
0	1	0	0.60	---	0	0
0	0	1	0.63	---	0	0
0.7	0.15	0.15	0.47	0.037	50	25
0.15	0.7	0.15	0.56	0.024	0	5
0.15	0.15	0.7	0.57	0.030	0	0
0.333	0.333	0.334	0.53	0.033	25	5

In general, the population predictions for Model 2 were higher than those for Model 1. This was not surprising, since in both models density was proportional to the amount of preferred habitat, with larger constants of proportionality in Model 2. For all combinations of state variables, the expected utility for Model 2 was equal to or greater than the expected utility for Model 1. Similarly, expected utilities for

Model 3 were generally equal to or greater than those for Model 2. As habitat was lost, Model 2 predicted immediate population equilibrium with the new conditions, while Model 3 predicted a time lag.

The optimal decisions also varied among the models. In general, the optimal amount to conserve was larger for the model which predicted lower population size.

Varying the likelihoods  $p_i$  assigned to each model affected both the optimal decision and the value of the objective function for the optimal decision (Table 2).

## DISCUSSION

The degree of confidence placed in each model affected both the optimal management decision and the expected value of the objective function. Basing management on an incorrect model without considering structural uncertainty can lead to suboptimal decisions.

The EVPI refers to the expected increase in the objective function from a hypothetical study designed to eliminate structural uncertainty. For example, a gain in objective function units of 0.033 might translate into an increase in the expected number of species conserved of  $0.033/W_1 = 0.176$ , or an increase in the proportion of land open for development of  $0.033/W_2 = 0.132$ . In a realistic management scenario, data will not discriminate among the models exactly. In that case, the expected gain from gathering the data depends on the model likelihoods (Lindley 1985). The expected value of partial information can be used in an analysis in the cost-benefit tradeoff of various levels of monitoring intensity.

In this exercise, we chose the model weights arbitrarily. An adaptive approach to management would involve treating the model weights as state variables whose dynamics are based on Bayesian updating. In passive adaptive management (Walters and Hilborn 1978), information feedback into the decision making process is not considered in evaluating the objective function. Optimization, management, and monitoring are done iteratively, with model weights recalculated after each iteration. In active adaptive management, the optimization specifically considers the benefits of future learning. Active adaptive management presents considerable computational difficulties, because the information state  $Y_t$  increases in dimension through time. A dynamic programming approach incorporating future learning must rely on a set of summary statistics which evolve in a Markovian fashion.

This model exercise represents a preliminary step in developing adaptive approaches to conservation planning. Our immediate objectives include exploring the effect of partial observability on decision-making in our model landscape, and performing similar analyses using population models with spatial structure. This will require some refinements in our solution methods due to the dimensionality limitations of dynamic programming. We anticipate taking an iterative simulation-optimization (SO) approach to dealing with more complex systems. In SO, methods such as dynamic programming or linear programming is used to develop optimal

policies for simplified or static systems. More detailed simulation models are then used to track the system dynamics over time using the optimal policy. If SO proves to be an effective strategy for dynamic optimization in complex systems, we will have the analytical tools available for addressing optimization that is truly adaptive, in the sense that the information state is included in the optimization procedure.

## ACKNOWLEDGMENTS

We thank Ken Williams, Jim Nichols, Fred Johnson, and Bill Kendall for invaluable consultation and input.

## REFERENCES

- Burley, F.W. 1988. Monitoring biological diversity for setting conservation priorities. pp. 227-230 in Wilson, E.O. (ed.) Biodiversity. Washington, DC:National Academy Press.
- Conroy, M.J. and B.R. Noon. Mapping of species richness for conservation of biological diversity: conceptual and methodological issues. Ecological Applications (in press).
- Crocker, J.E. and M.J. Conroy. 1995a. A framework for optimal decision making in the management of landscapes. 80th Annual Meet., Ecological Society of America, Snowbird, Utah. 30 July-3 August 1995. (Abstract).
- Keeney, R.L., and H. Raiffa. 1976. Decisions with multiple objectives. New York:Wiley.
- Lindley, D. V. 1985. Making decisions. New York:Wiley 207 pp..
- Lubow, B. 1993. Stochastic Dynamic Programming (SDP): User's Guide. Fort Collins, CO :Colorado Coop. Fish and Wildlife Res. Unit.
- Nichols, J.D., F.A. Johnson, and B.K. Williams. 1995. Managing North American waterfowl in the face of uncertainty. Annual Review of Ecology and Systematics 26:177-199.
- Pulliam, H. R. 1988. Sources, sinks, and population regulation. American Naturalist 132:652-661.
- Scott, J. M., F. Davis, B. Csuti, R. Noss, B. Butterfield, C. Groves, H. Anderson, S. Caicco, F. D'Erchia, T. C. Jr. Edwards, J. Ulliman, and R. G. Wright. 1993. Gap analysis: a geographic approach to protection of biological diversity. Wildlife Monographs 123:41pp.
- Walters, C.J. 1986. Adaptive management of renewable resources. New York: MacMillan. 374 pp.
- Walters, C.J., and R. Hilborn. 1978. Ecological optimization and adaptive management. Annual Review of Ecology and Systematics 9:157-188.
- Williams, B.K. 1989. Review of dynamic optimization methods in renewable resource management. Natural Resource Modeling 3:137-216.

# Sensitivity Analysis of Species Richness Mapping to Variations in Forest-Wildlife Relationships

Thomas W. Kohley<sup>1</sup>

---

**Abstract.** The Gap Analysis Program, administered by the National Biological Service, was implemented to identify "gaps" in the conservation of selected surrogates of biodiversity. Information on habitat affinities and the known occurrence of terrestrial vertebrate species are incorporated into a GIS to develop maps of predicted species distributions and richness. With the incorporation of species-habitat suitability rankings into the Gap models, it is possible to geographically delineate areas of critical, acceptable, and marginally suitable habitats for each species. Further, species richness can be mapped using a "liberal" or "conservative" approach. The liberal approach taken by Gap Analysis contends that all habitats, whether high or low value for a species, are of conservation value and should be included in the measurement of species richness. The conservative approach argues that only the critical habitats warrant the highest conservation priority and all other habitat-associations should be omitted from estimates of species richness. This research examined the sensitivity of species richness to these two mapping approaches and explored the potential effects on biological conservation planning. Results indicated that the conservative approach differs substantially from the liberal approach in terms of (1) the predicted representation of animal species within existing reserves and (2) the identification of species rich hotspots.

---

---

<sup>1</sup> GIS Research Scientist, Wyoming Water Resources Center, Laramie, WY.

# Development of a Survey Sampling Methodology for Rare Species

Molly Van Caster\*    David Bowden, Ph.D.<sup>†</sup>  
Jennifer Hoeting, Ph.D.<sup>‡</sup>

---

**Abstract.**—Developing a plan to find rare species is, by nature, a difficult task. Incorporating habitat information can improve sampling plans. In our current application, the available habitat information is rather sparse which increases the difficulty in developing a useful sampling plan.

The information comes from past sightings, if they occurred, with or without habitat information, scientists' knowledge, if available, and information on the study area in the form of GIS and from survey plots. The key is to link the habitat information available to possible sampling locations.

The model developed predicts likelihood of presence at a specific site. The information from the site is used, as well as habitat information from the neighboring sites. Neighboring sites are those which have presence/absence data along with some habitat information. Thus, sites similar to sighting locations would be deemed more likely to support the species. The model and simulation studies will be presented.

---

---

\*Ph.D. Candidate, Department of Statistics, Colorado State University, Fort Collins, CO.

<sup>†</sup>Professor, Department of Statistics, Colorado State University, Fort Collins, CO.

<sup>‡</sup>Assistant Professor, Department of Statistics, Colorado State University, Fort Collins, CO.

# Methods to Analyse the Spatial Structure of Plant Communities

Paul Braun, Heiko Balzter and Wolfgang Köhler<sup>1</sup>

---

**Abstract.** - Though many methods for analysing the spatial structure of plant communities are available, few of them deal with the problem of clonal plant growth or spatio-temporal dynamics. This paper shows two possibilities for quantifying the spatio-temporal dynamics of a plant population and a vegetation. It turns out, that fine shifts in plant populations or plant community structure can be discovered, if spatial and temporal dynamics are analysed simultaneously. Still there remains the need for improved or new methods.

---

## INTRODUCTION

Any knowledge about the spatial distribution of a plant population allows to draw conclusions on predominant biotic or abiotic factors. This in turn enables us to better predict the population dynamics of a specific plant population or even a vegetation.

If discrete objects/individual plants are identifiable, common methods of quantitative ecology are used to further process the data (Greig-Smith 1983, Elliott 1971, Kershaw 1978, Upton and Fingleton 1990, Vandermeer 1981). These methods can be grouped as either aggregation or distance methods. While aggregation methods count individuals (e.g. Variance to Mean Ratio, Morisita Index), distance methods (e.g. Nearest Neighbour Index, L-Plot) utilize inter-plant distances to conclude, whether plants are randomly, clustered or regularly dispersed. When plants no longer can be assigned to one point, as is the case with clonal plants or vegetations, the above mentioned methods are less useful to describe spatial dynamics of populations.

In vegetation science different procedures are used to quantify the proportions of species in a plant community, where discrete objects cannot be counted, e.g. Braun-Blanquet-, Quadrat-, Point-Quadrat-Method. Depending on the type of data, various methods can be applied to detect spatial structures. The Point-Quadrat-Method provides a good base to apply a Moving Window Algorithm (Balzter et al. 1995).

---

<sup>1</sup>*Justus-Liebig-University, Dep. of Biometry and Population Genetics, Ludwigstr. 27, D-35390 Giessen, Federal Republic of Germany; email gh80@agrar.uni-giessen.de*

**THE SPATIO-TEMPORAL DYNAMICS OF A THISTLE-POPULATION  
(*CIRSIUM ARVENSE* (L.) SCOPOLI).**

After mapping plant populations on a recultivated area close to Giessen in 1991, we used these data to examine the usefulness of various aggregation and distance methods to quantify the spatial distribution of plant populations (Braun and Lachnit 1993, Braun and Lachnit 1994). We found the Nearest Neighbour Index ( $R$ ) to be a suitable statistic for describing spatial and temporal changes in plant populations of discrete objects.  $R$  depends on data about the distance of each plant to its nearest neighbour. As under field conditions this information may not easily be obtained, we mapped the whole plant population (*C. arvense*) in a 2x2 m plot. With these coordinates the average distance,  $r_M$ , of all individual plants to its nearest neighbour had been calculated and normalized by the expected nearest neighbour distance,  $r_E$ . This is expressed formally by

$$R = r_M / r_E \quad [1]$$

where  $R$  is the ratio between the measured average nearest neighbour distance  $r_M$  and the expected nearest neighbour distance  $r_E$  (i.e. random distribution of individual plants in the population). The value of  $r_M$  is obtained by

$$r_M = (1/N) \cdot \sum_{i=1}^N r_i \quad [2].$$

where  $r_i$  is the distance of plant  $i$  to its nearest neighbour and  $N$  is the population size. The expected value  $r_E$  is defined as

$$r_E = 1 / (2 \cdot \sqrt{u}) \quad \text{with } u = N/A \quad [3].$$

Here  $A$  stands for the area of the mapped plot.  $R$  close to 1 indicates randomness,  $R < 1$  indicates a clustered and  $R > 1$  indicates a regularly dispersed plant population. To test the significance of  $R$ , Clark and Evans (1954) proposed the following test statistic:

$$CE = R / \sigma_E, \quad \text{where} \quad [4]$$

$$\sigma_E^2 = (4 - \pi) / (4 \cdot N \cdot u \cdot \pi) \quad [5].$$

CE is standardnormally distributed and can be tested by z-values. For six different sampling dates in 1991 we calculated  $R$ . In Figure 1 the dynamics of the  $R$  values for the thistle-population plot is shown.

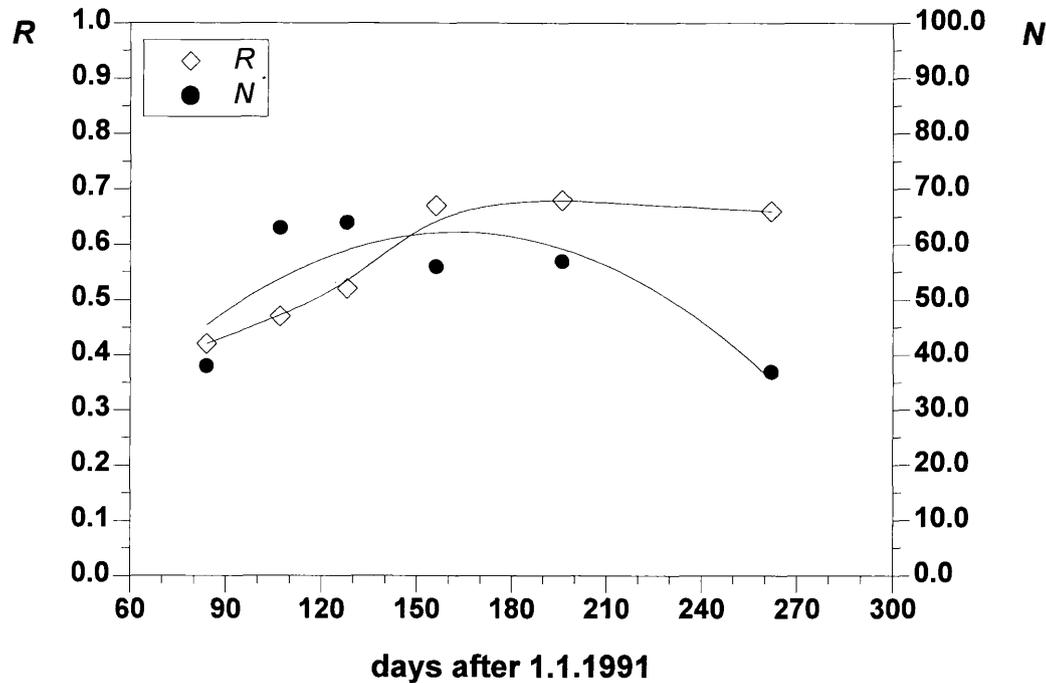


Figure 1. - Spatial (Nearest Neighbour Index) and temporal (population size) dynamics of a thistle-population in 1991. Sampling dates were: 25.3., 17.4., 8.5., 5.6., 15.7. and 19.9.1991.

The population size (measured by  $N$ ) stays steady except for a temporary seasonal increase. This population dynamics is quite usual for natural populations in central Europe. Our data suggest a decrease in clustering of the thistle-population during the year. It seems that the thistle-population spreads, although it does not increase in number. This could be due to competition, because grouped thistle plants experienced stronger competition than single plants. The observed loss in aggregation would not have been detected by only looking at the population dynamics.

### THE SPATIO-TEMPORAL DYNAMICS OF A MEADOW VEGETATION (*Lolio-Cynosuretum*).

In the continuous case, i.e. when plants are not viewed as single points, aggregation and distance methods are not efficient. We therefore recorded an area by the Point-Quadrat-Method (PQM) and processed these raw data with a Moving Window Algorithm (MWA).

Our investigation was conducted on a meadow in Giessen. Except for the centre, which remained unmown all the time, the meadow was cut up to ten times a year.

This led to an association of *Lolio-Cynosuretum* on the outer part of the meadow (Figure 2). Whereas in the unmown centre only a class of *Molinio-Arrhenatheretea* could be observed (Oberdorfer 1990).

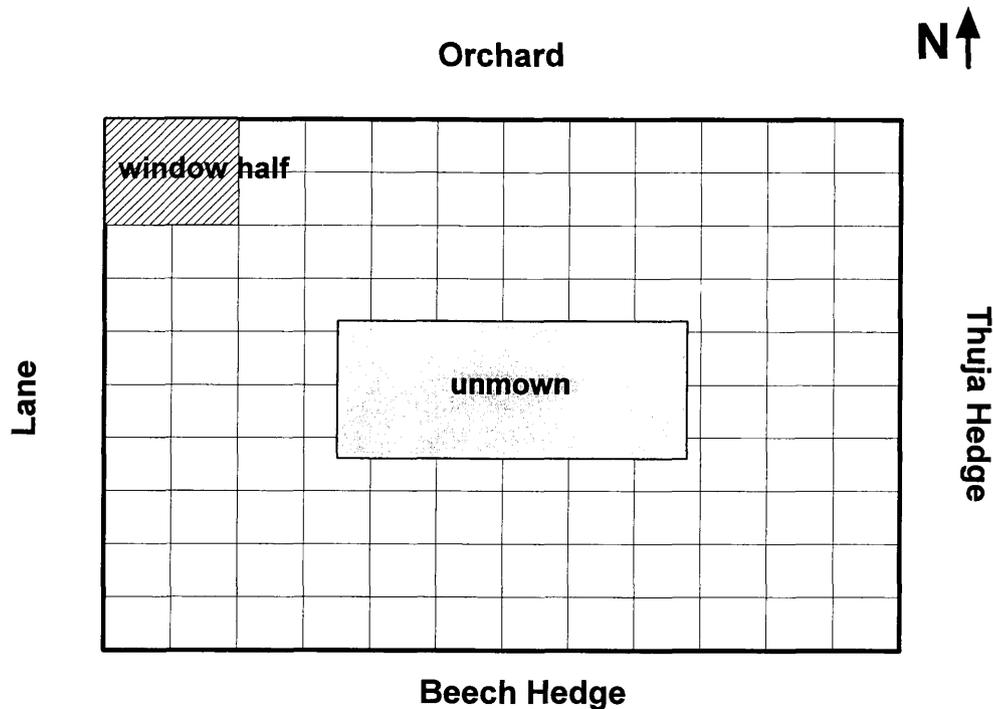


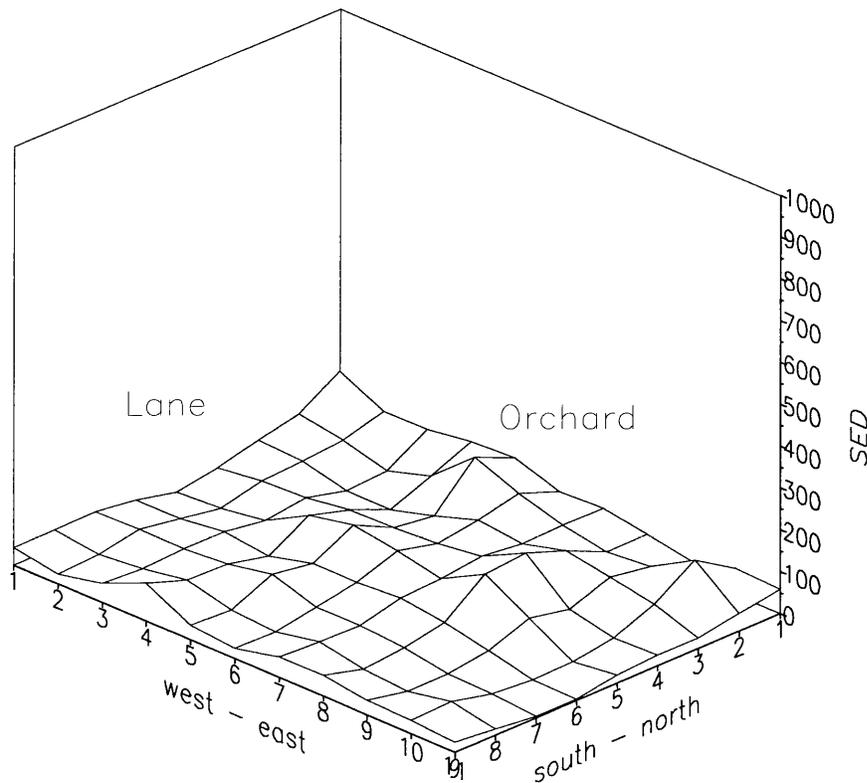
Figure 2. - Investigated meadow in Giessen (630 m<sup>2</sup>). The whole area is divided in 120 subplots, which were sampled every three months. The central part remained unmown (72 m<sup>2</sup>). The scientific name for thuja is *Thuja occidentalis* and for beech *Carpinus betulus*.

As shown in Figure 2 the whole area was divided into 120 subplots. Inside of each subplot a vegetation sample was taken with a point quadrat frame (three needles, ø 1 mm, each 30 cm apart). A vegetation sample means, that all contacts of plant species with any of the three needles in one frame were recorded.

Four subplots or one window half yielded a multivariate vector, where each variable contained the mean number of contacts that one plant species had with any of the needles. To compare two successive years, the vegetation data of the same window half, but for two successive years were used. These two window halves - a window - represented two multivariate vectors for which the Squared Euclidian Distance (SED) was calculated:

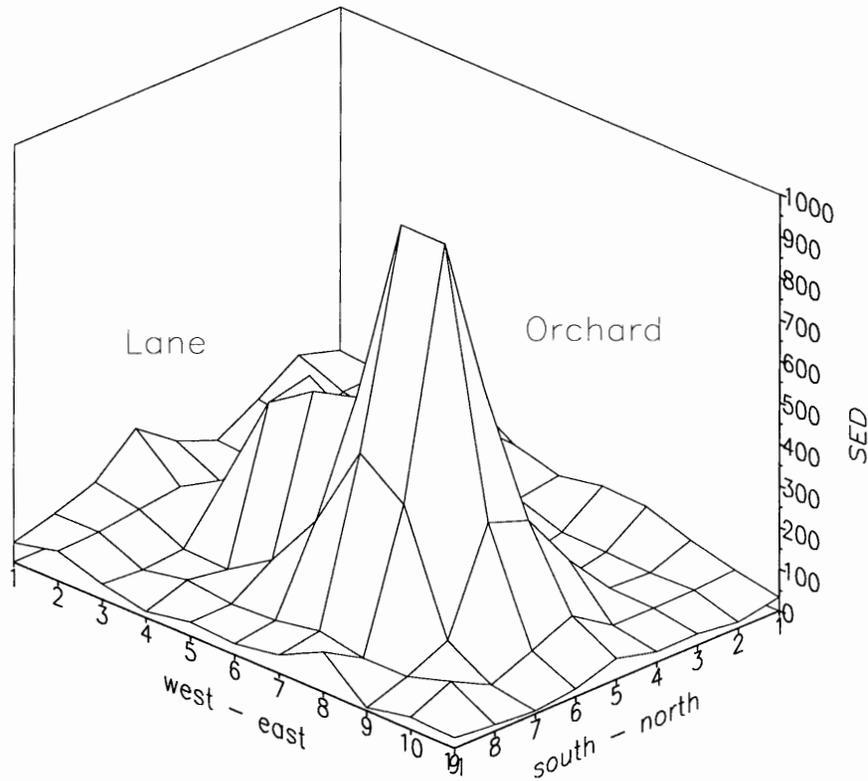
$$SED_{mw} = \sum_{i=1}^s (\bar{x}_{Ai} - \bar{x}_{Bi})^2 \quad [6].$$

Variables  $n$  and  $w$  denote the number and size of the window, respectively. The total species number is  $s$ . Index  $A$  is the window half of the first year and  $B$  the window half of the second year. An SED value of zero would mean that no vegetation change has occurred. High SED values indicate large vegetation change between the years. In our example we moved the window stepwise, that means one subplot, from left to right (Fig. 2). Thus the window stretched over two years but was moved in space. For one row this gave 11 values. After finishing one row the procedure continued with the next row. Thus the vegetation change between years is described by 11x9 points, where each value represents a window comparing two years. Figure 3 shows the vegetation change from 1993 to 1994.



**Figure 3. - Plot of the SED's between the vegetation from 14.5.1993 and 20.6.1994. In the background the orchard is located, on the left hand stretches the lane, etc. (see Fig. 2).**

There are fluctuations in the vegetation composition, but there is no obvious change in vegetation. To get an idea on the importance of peaks, we compared the SED values of 1993/94 (Fig. 3) with the corresponding values for 1994/95. (Figure 4).



**Figure 4. - Plot of the SED's between the vegetation from 20.6.1994 and 16.6.1995. In the background the orchard is located, on the left hand stretches the lane, etc. (see Fig. 2).**

This comparison revealed a substantial change for the period 1994/95. Indeed even this big shift was not easy to detect by a routine inspection out in the field. After we knew these results we searched for the changes on the meadow and found them. But this shift might have easily been overlooked, without an MWA-analysis. This method has proven to be a very sensitive way to detect vegetational changes.

We assume that this phenomenon was the result of massive herbivory by a local rabbit population which augmented considerably during the mild winters of 1993/94 and 1994/95 in Giessen.

## CONCLUSIONS

Our experience with methods describing spatio-temporal dynamics leads us to the conclusion that there are only a few efficient methods to solve this task. As Ripley (1986) pointed out most of the aggregation methods and some of the distance methods can be misleading in the interpretation of spatial structures. For a

close inspection of a spatial structure Ripley's L-Plot is recommended. But the information density of this statistic made us prefer  $R$ , which easily can be combined with population dynamics data in one plot. Even small changes in structure are detectable thus. Still there is one problem unresolved: the form and position of structures in an investigation plot cannot be described. This information might be necessary to judge specific population dynamics based on clustered resources. Because of this, but not only, we recommend to always take a look on the raw data.

The combination of PQM and MWA gives good estimates of even slight vegetation, population or biomass changes. But one has to keep in mind, that the SED values are relative values. This means we can only compare the distance between specific window halves. In other words, two SED's that are plotted aside in Figure 3 or 4 may derive from different species compositions, even if they have the same value. Another lack exists for the MWA: for our problem we did not find a standard statistical test of significance.

Although there are still some methodological improvements needed, the study of spatio-temporal dynamics of herbaceous plant populations should be intensified.

## REFERENCES

- Balzter, H. , Braun, P.W. and Köhler, W. 1995: Detection of spatial discontinuities in vegetation data by a moving window algorithm. In: From Data to Knowledge: Theoretical and Practical Aspects of Classification, Data Analysis and Knowledge Organization. (Eds.: Gaul W. and Pfeifer D.). - Springer, Heidelberg: 243-252.
- Braun, P. and Lachnit, B. 1993: Characterizing the spatial distribution categories in plant populations. - Biometrical Bulletin 10(3): 25.
- Braun, P. and Lachnit, B. 1994: Kennzeichnung der räumlichen Verteilung von Pflanzenpopulationen. - Zeitschrift für Agrarinformatik 4: 67-71.
- Clark, P.J. and Evans, F.E. 1954: Distance to nearest neighbour as a measure of spatial relationships in populations. - Ecology 35: 445-453.
- Elliott, J.M. 1971: Some methods for the statistical analysis of samples of benthic invertebrates. - Freshwater Biological Association, Ambleside. 144 p.
- Greig-Smith, P. 1983: Quantitative Plant Ecology. - Blackwell Scientific Publishers, Oxford.
- Kershaw, K.A. 1978: Quantitative and Dynamic Plant Ecology. - Edward Arnold, London.
- Oberdorfer, E. 1990: Pflanzensoziologische Exkursionsflora. - Ulmer, Stuttgart.

- Ripley, B.D. 1986: Spatial point pattern analysis in ecology. In: Developments in Numerical Ecology. (Eds.: Legendre, P. and Legendre, L.). - Springer, Berlin: 407-429.
- Upton, G.J.G. and Fingleton B. 1990: Spatial Data Analysis by Example. - Wiley, Chichester/New York.
- Vandermeer, J. 1981: Elementary Mathematical Ecology. - Wiley, New York.

### **BIOGRAPHICAL SKETCH**

Paul Braun is agricultural scientist in the Department of Biometry and Population Genetics in Giessen. He obtained a Dr. agr. degree at Giessen University in 1988 and works on agroecology and environmental informatics.

Heiko Balzter is PhD student in the Department of Biometry and Population Genetics in Giessen. He holds a Dipl.-Ing agr. degree from Giessen University since 1994 and works on spatio-temporal modeling in landscape ecology.

Wolfgang Köhler is Biometrician and Professor in the Department of Biometry and Population Genetics in Giessen since 1980. His research mainly concerns population genetics, environmental modeling, experimental design and statistical evaluation of field trials.

# Modeling Population Dynamics with Cellular Automata

Heiko Balzter, Paul Braun and Wolfgang Köhler<sup>1</sup>

---

**Abstract:** When predicting changes in plant populations, spatial interactions must be considered. Cellular automata are models incorporating both spatial and temporal dynamics. Vegetation data were sampled on a meadow over three years using the point-quadrat method and population dynamics of three single species were modeled with a cellular automaton using stochastic decision rules. Links with the theory of Markov chains are briefly discussed.

---

## INTRODUCTION

Ecological science is structured in the three major disciplines autecology, demecology and synecology. While autecology examines the relation between a single organism and its environment, demecology looks at a whole population of organisms and synecology investigates several populations of different species interacting with the environment. In this sense our approach to modeling spatio-temporal population dynamics is an application of statistical methods to a synecological problem. Because there is still a lack of a complete theory of succession (Huston 1994), the need for building vegetation models to simulate changes and patterns becomes obvious. Cellular automata provide a class of spatio-temporal models with a simple basic structure but offer a nearly unlimited range of possibilities. Their value for modeling purposes is discussed, looking at a vegetation study performed at the Agricultural School of Giessen University, Germany, from 1993 till today.

## CELLULAR AUTOMATA

A cellular automaton model consists of  $N$  elements, or cells, each capable of several discrete values in a defined state space. The cells can be linked in different ways. In the simplest case they are connected geometrically according to any spatial order, such as in a one- or two-dimensional grid, in which the spatially neighboring cells determine the state of each cell. But the cells may be connected randomly, too, such as in neural networks (Kauffman 1984). Here, only simple,

---

<sup>1</sup> Address: Justus-Liebig-University, Dep. of Biometry and Population Genetics, Ludwigstrasse 27, D-35390 Giessen, Federal Republic of Germany. Email: gha9@agrar.uni-giessen.de

two-dimensional automata with spatially symmetrical neighborhood links are considered. Neighborhood definitions often used are the Moore-neighborhood consisting of the cell itself and the eight neighboring cells with at least one point in common, and the von Neumann-neighborhood, that only contains the cell itself and the four cells with at least one side in common. Each cell can change its state during discrete time steps  $t$ , according to specified decision rules<sup>2</sup>, which can be either deterministic or stochastic (Czárán & Bartha 1992).

For a step towards a theory of cellular automata see Wolfram (1984). Cellular automata have a broad range of applications, for example to model chemical reactions with spatial diffusion, the development of spiral galaxies, for phase transitions, crystal growth, but also quite often to model biological and ecological systems (Wolfram 1983, Molofsky 1994). A literature review concerning ecology is given by Phipps (1992). Looking exclusively at the modeling approaches to ecological systems one furthermore must emphasize the following publications.

Flamm & Turner (1994) modeled land cover with a stochastic cellular automaton integrating several data layers and thus increasing precision of the model output. In vegetation science, Ratz (1994) made forecasts of the development of spatial structures of boreal forests underlying the influence of fire. The effect of fire on *Banksia* populations in Australia was also examined (Marsula & Ratz 1994). On a small scale, Winkler et al. (1994) modeled the dynamics of a dry grassland community, observing a plot sized one square meter. A very important study was performed by Silvertown et al. 1992, who modeled population dynamics of five grass species, paying special attention to competition. They extended the Markovian property to a cellular automaton model and obtained results that a non-spatial model would not have been able to explain. A study of the impacts of a coal-fired power plant to freshwater wetland in Wisconsin was carried out by Ellison & Bedford (1995) with a quite sophisticated model. Using rank data the model output agreed well with the observed changes caused by hydrological impact. A major problem of cellular automata models is to formulate adequate decision rules determining the state transitions of the cells. A good solution to this is presented in Wiegand et al. (1994), who modeled five plant species of a semi-arid shrub ecosystem in South Africa, translating biological and ecological knowledge into decision rules depending on weather conditions. Their major problem was the time scale of the model, as approximately every ten years the seemingly stable vegetation abruptly changed, following a certain weather event. However, it is possible to successively improve the decision rules by the "top-down approach" as described by Kummer et al. (1994). Modeling the spread of rabies they started with a simple, one-dimensional basic model and scaled down in the investigation of the system by expanding the model spatially to two dimensions with time steps of one year, and then temporally to finer time steps of two months. Their experiences were satisfying, for it was possible this way to gain heuristic insights in the dynamics of the rabies-fox-system.

---

<sup>2</sup> sometimes called "transition rules" or simply "rules"

In a detailed discussion of cellular automata and ecological theory, Phipps (1992) comes to the conclusion that, as far as application to natural systems is concerned, probabilistic decision rules usually have a better analogy to the system than deterministic ones, though sometimes their heuristic value is less.

Kareiva & Wennergren (1995) pointed out that a strength of this class of models is that it is capable of modeling ecosystems with respect to spatial segregation, which often is a condition for coexistence of predator and prey populations. Possibly cellular automata can contribute ideas to solving the problem of competition and coexistence of plant species (Grace 1995).

## MATERIAL AND METHODS

The meadow is located in Giessen, Germany. It is regularly mown about ten times a year apart from an area in its centre. The mown plant community was classified as *lolio-cynosuretum* following from the phytosociological system of Braun-Blanquet (1964), whereas the unmown plant community had a quite different composition. For reason of simplicity this difference in use was ignored in our model. The survey method used was the point-quadrat method. An extensive statistical study of point-quadrat methods was performed by Goodall (1952). The frame used in this study holds three pins above the vegetation, and during sampling each pin passes through a guide channel down to the ground and the number of contacts of the pins to each plant species is counted (figure 1). In this manner a grid of 10 parallel transects, each consisting of 12 frame positions was sampled, resulting in 120 subplots. Methodological issues are discussed elsewhere (Balzter et al. 1995). Data sets for *Lolium perenne* (perennial ryegrass), *Trifolium repens* (white clover) and *Glechoma hederacea* (ground ivy) from May 1993, June 1994 and June 1995 were used.

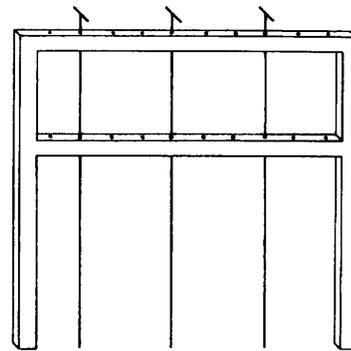


Figure 1.- Point-quadrat frame as described by Kreeb (1983) and used in this study.

The 120 subplots on the meadow are used as the cells of the two-dimensional cellular automaton. When looking at a certain plant species, it can either be absent (=0) or present (=1) in each cell, resulting in a dualistic, discrete state space. First, the transitions from 1993 to 1994 and from 1994 to 1995 were pooled. In order to set adequate decision rules, two different approaches to stochastic probability estimation were then carried out:

- 1) Under the assumption that only the state of the cell itself at the preceding time step determines its actual state, a matrix of transition probabilities (transition matrix) as known from the theory of Markov chains was estimated from the data.

Thus, the „neighborhood“ incorporates only the subplot itself. We will refer to this approach as „spatial Markov chain“.

2) Assuming an influence of the spatially neighboring subplots on the subplot itself, the Moore-neighborhood (the cell itself and the eight neighboring cells) was used in the second set of decision rules. Transitions from any possible neighborhood into one of the two states must be considered. Because this would result in 36 different neighborhood states, they were classified, so that the transition matrix is stochastic<sup>3</sup>. Obviously this classification has to be changed for frequent and rare species (table 1). In the following text this approach is called „Moore-neighborhood“.

**Table 1.- Neighborhood classifications. Numbers mean neighboring cells with species present. Percentage cover was smallest for *Glechoma hederacea* , followed by *Lolium perenne* and finally *Trifolium repens*.**

neighborhood class	<i>Lolium perenne</i>	<i>Trifolium repens</i>	<i>Glechoma hederacea</i>
0	0	0	0
1	1-2	1-2	1
2	3-4	3-4	2
3	5-7	5-7	3
4	8-9	8-9	4-9

Once the transition matrix was estimated, the distributions of cells with absence or presence of the species at several time steps were predicted. During each time step every cell randomly took a new state depending on the corresponding probability distribution of the transition matrix.

## RESULTS

The transition matrices for the spatial Markov chain estimated from the vegetation data are given in table 2.

**Table 2.- Transition matrices for the spatial Markov chain and three plant species.**

from	to	<i>Lolium perenne</i>		<i>Trifolium repens</i>		<i>Glechoma hederacea</i>	
		0	1	0	1	0	1
0 (absent)		(0.55 0.45)		(0.53 0.47)		(0.92 0.08)	
1 (present)		(0.45 0.55)		(0.23 0.77)		(0.62 0.38)	

For the Moore-neighborhood the matrices (table 3) must not be confused with the transition matrices of Markov processes, because the neighborhood state space is not the same as the cell state space, expressed in the different numbers of rows

<sup>3</sup> A matrix is called „stochastic“, if i) all row sums equal one, ii) no element is less than zero or greater than one and iii) at least one element in each column differs from zero.

and columns of the matrices. Because of this property is it far from easy to interpret the matrices in the way we do with the spatial Markov chain.

**Table 3.- Transition matrices for the Moore-neighborhood using five neighborhood classes for the three plant species.**

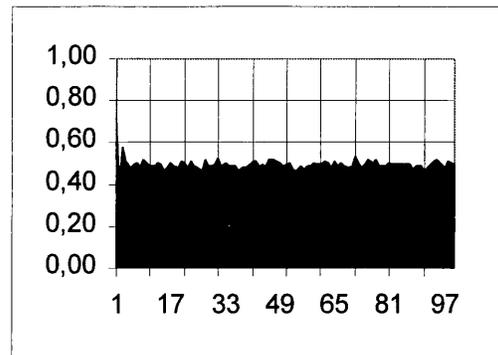
from neighborhood class	to	<i>Lolium perenne</i>	<i>Trifolium repens</i>	<i>Glechoma hederacea</i>
		0 1	0 1	0 1
0		(0.91 0.09)	(1.00 0.00)	(0.99 0.01)
1		0.52 0.48	0.54 0.46	0.73 0.27
2		0.54 0.46	0.50 0.50	0.86 0.14
3		0.46 0.54	0.27 0.73	0.29 0.71
4		(0.41 0.59)	(0.18 0.82)	(0.33 0.67)

Figure 2 shows the dynamics of *Lolium perenne*. The relative frequency of cells with the species present quickly reaches a limiting distribution and the following variation is caused by the limitation of the number of cells. If an infinite number of objects were to pass the Markov chain, the limiting distribution would be stable. Interested in the value of the limiting distribution we can calculate a first estimate from the last 20 time steps  $t = 81, 82, \dots, 100$ . For *Lolium perenne* this results in  $\bar{x}_{Lp} = 0.50$ .

For *Trifolium repens* the predicted frequencies show similar results, but stabilize on a higher level ( $\bar{x}_{Tr} = 0.67$ , figure 3) and *Glechoma hederacea* tends against  $\bar{x}_{Gh} = 0.12$  (figure 4).

The theory of Markov chains allows us to calculate the limiting distributions, assuming that every subplot is changing its state according to the Markovian transition matrix. Because all three transition matrices  $P$  of the spatial Markov chains are aperiodic classes of positively recurrent states and the state space is finite, the limiting distributions  $p(\infty)$  are ergodic and can be determined

by eq. 1 (Heller et al. 1978). Note, that the ergodic property means, that the limiting distribution is independent of the initial distribution. In other words, the plant population reaches the same equilibrium, wherever it starts. The results (table 4) agree to the rough estimates from the data.



**Figure 2.- Relative frequency of cells colonized by *Lolium perenne* over 100 time steps (mean of 10 simulation runs) predicted by the spatial Markov chain. x:time; y: relative frequency.**

$$\underline{p}(\infty) = \lim_{t \rightarrow \infty} P^t \quad [1]$$

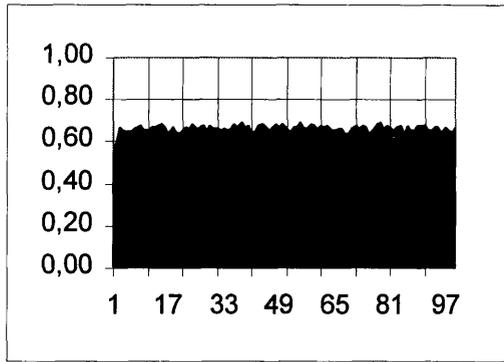


Figure 3.- Relative frequency of cells colonized by *Trifolium repens* over 100 time steps (mean of 10 simulation runs) predicted by the spatial Markov chain. x:time; y: relative frequency.

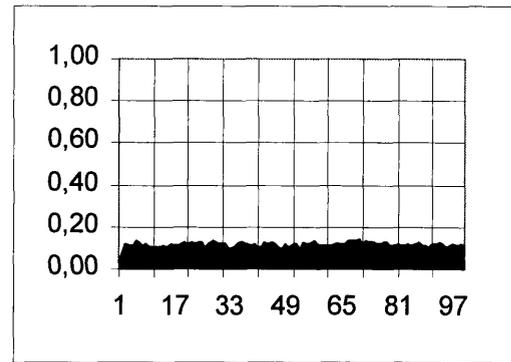


Figure 4.- Relative frequency of cells colonized by *Glechoma hederacea* over 100 time steps (mean of 10 simulation runs) predicted by the spatial Markov chain. x:time; y: relative frequency.

Table 4.- Ergodic distributions of the spatial Markov processes

	<i>Lolium perenne</i>		<i>Trifolium repens</i>		<i>Glechoma hederacea</i>	
state:	0 (absent)	1 (present)	0 (absent)	1 (present)	0 (absent)	1 (present)
rel. freq.:	0.50	0.50	0.33	0.67	0.89	0.11

Using the second set of decision rules, under the assumption of the Moore-neighborhood determining the future state of a cell, the model still shows asymptotical behavior. Figure 5(a) shows the results for *Lolium perenne*. Surprisingly they do not differ from those of the spatial Markov chain, as the estimated limiting distribution  $\bar{x}_{Lp} = 0.50$  is the same. To examine the model behavior two further approaches were made. First the transition matrix was applied to various initial distributions. All initial distributions tended to vary randomly around the same mean after a specific number of time steps. Secondly, the cell grid of the cellular automaton was expanded to  $100 \cdot 100 = 10000$  cells.

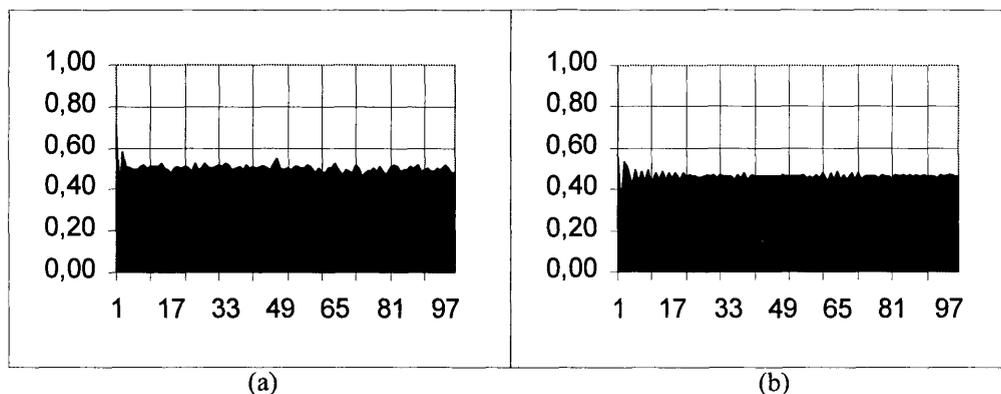


Figure 5.- Relative frequency of cells colonized by *Lolium perenne* over 100 time steps as predicted by the cellular automaton using the Moore-neighborhood. x:time; y: relative frequency. (a) mean of 10 simulation runs; (b) one simulation run, 100x100 cells.

As expected the variation was decreased strongly by this expansion (figure 5(b)), while the mean of the limiting distribution is shown more clearly than using 120 cells.

Looking at *Trifolium repens* the results are similar (figure 6), but setting the complete initial distribution to zero no plant can ever establish itself. The reason for this is the transition matrix (table 3), where a neighborhood of all zero produces a cell in state zero with probability  $p_{00} = 1$ . The limiting distribution of *Trifolium repens* is thus dependent on the initial distribution. Neglecting this special case, because for practical purposes it is very unlikely that no single plant of *Trifolium repens* would be found on the meadow, the estimated limiting distribution is  $\bar{x}_{Tr} = 0.72$  differing slightly from the forecast of the spatial Markov chain.

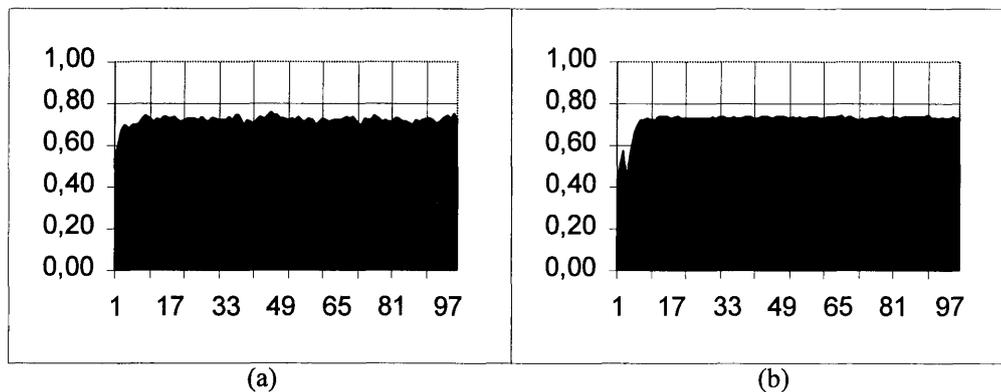


Figure 6.- Relative frequency of cells colonized by *Trifolium repens* over 100 time steps predicted by the cellular automaton using the Moore-neighborhood. x:time; y: relative frequency. (a) mean of 10 simulation runs; (b) one simulation run, 100x100 cells

A completely different result is obtained for *Glechoma hederacea* in figure 7. Although the model behavior is still the same as for the other species, the value of the limiting distribution  $\bar{x}_{Gh} = 0.66$  is much greater than that given by the spatial Markov chain. Which model comes closer to reality can hardly be judged, but will be subject to further validation in the following years. Because *Glechoma hederacea* does not typically cover two thirds of an area the spatial Markov chain is expected to be more reliable, but this is only supposition. The importance of selecting appropriate decision rules based on the right neighborhood definition can be seen in this example. If the neighboring subplots do not have a major impact on the species in the subplot, the Moore-neighborhood will be the wrong modeling approach and a spatial Markov chain will be preferred. But, if the species is able to colonize adjacent subplots by stolons or seed dispersal, the Moore-neighborhood is likely to be preferable. Using stochastic decision rules whose probabilities are estimated from the data has the major advantage that the system behavior can be modeled without knowing exactly quantitative relationships between certain factors (which also would result in a quite complicated

deterministic model). The effects of all major factors influencing the development of the population add up to a stochastic probability density function that can easily be determined.

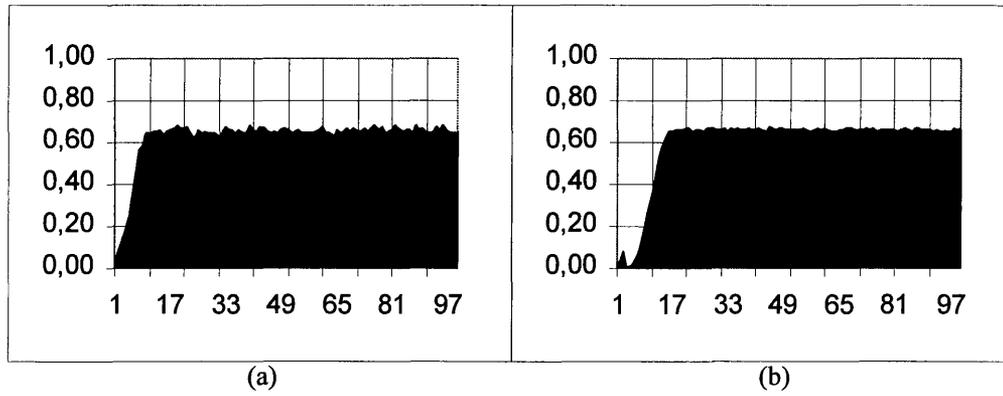


Figure 7.- Relative frequency of cells colonized by *Glechoma hederacea* over 100 time steps predicted by the cellular automaton using the Moore-neighborhood. x:time; y: relative frequency. (a) mean of 10 simulation runs; (b) one simulation run, 100x100 cells

## DISCUSSION

The theory of cellular automata has not yet been satisfactorily investigated. Despite the underlying simplicity of this class of models, cellular automata exhibit numerous different behaviors. The most important work was done by Wolfram (1984), but there is still a need for mathematical examinations. In our study the model behavior can be explained partly using the theory of Markov chains. In addition, the proved existence of a limiting distribution is important for the assumptions of vegetation science. A limiting distribution is equivalent to the concept of a climax of succession as proposed by Clements (1916). Whether there is a climax of the observed vegetation changes or not shall not be discussed in detail, because various conditions necessary for the maintenance of the successional path are likely to change before this question arises, e.g. the size of the rabbit population. These changing conditions erect limits for the predictions of the cellular automaton model. 100 time steps are surely too many to interpret in a serious way, this number was simply selected to examine the long-term model behavior. But the model output described above probably gives a rough impression of the expected development of cover and random variation of the plant species.

The two cellular automata presented here show differences in model output for two of the three species. Because the only thing that was changed was the definition of the stochastic decision rules, the differences in model output must be due to assumed spatial interactions between adjacent subplots. The way in which cells are connected thus plays an important role. Furthermore, the decision rules used here are very simple: They cover only one single species and do not contain

information about biological conditions of the species, such as mean length of stolons or distance of seed dispersal, which were included in the studies of Wiegand et al. (1994) and Ellison & Bedford (1995). In fact, it is intended to improve successively the decision rules of our model following the top-down approach of Kummer et al. (1994) by using vegetation data of four time steps per year, information about the seed bank in the soil, wind dispersal and interactions between species (the latter was done for two species by Balzter et al. 1996).

However, a practical advantage of strongly simplified decision rules in the model is that data sampling is not as laborious as if a lot of factors had to be measured. This feature is even more important, if the difficulty of sampling spatial data is considered. This may be one reason, why over a long time only a few scientists have applied spatial data analyses to their problems, although in the meantime the importance of both spatial and temporal structures and dynamics are generally recognized. In the future more spatial data will be usable, because Geographic Information Systems (GIS) and large data bases are becoming more and more established. Linking cellular automata with raster-based GIS seems to be a promising approach. The great freedom in setting up decision rules allows the adaptation of a wide range of knowledge to the model and prediction of the expected development under the condition that the assumptions made in the model are true. This variety of possibilities might promote the use of cellular automata in future ecology.

## REFERENCES

- Balzter, H. , Braun, P. & Köhler, W. 1995. Detection of Spatial Discontinuities in Vegetation Data by a Moving Window Algorithm. In: Gaul, W. & Pfeifer, D. (Eds.): From Data to Knowledge: Theoretical and Practical Aspects of Classification, Data Analysis and Knowledge Organization. Springer, Berlin, 243-252.
- Balzter, H., Braun, P.W. & Köhler, W. 1996 (in press). Ein zelluläres Automatenmodell zur Simulation der Vegetationsdynamik einer Weidelgras-Weißklee-Weide. Tagungsberichte der AG Ökologie der Biometrischen Gesellschaft, Freiburg & Leipzig.
- Braun-Blanquet, J. 1964. Pflanzensoziologie. Springer, Wien & New York.
- Clements, F.E. 1916. Plant Succession. Carnegie Institute of Washington Publication 242,1-512.
- Czárán, T. & Bartha, S. 1992. Spatiotemporal Dynamic Models of Plant Populations and Communities. Trends in Ecology and Evolution 7:2, 38-42.
- Ellison, A.M. & Bedford, B.L. 1995. Response of a Vascular Plant Community to Disturbance: A Simulation Study. Ecological Applications 5:1, 109-123.
- Flamm, R.O. & Turner, M.G. 1994. Alternative Model Formulations for a Stochastic Simulation of Landscape Change. Landscape Ecology 9:1, 37-46.

- Goodall, D.W. 1952. Some Considerations in the Use of Point Quadrats for the Analysis of Vegetation. *Australian Journal of Scientific Research*, B 5, 1-41.
- Grace, J.B. 1995. In Search of the Holy Grail: Explanations for the Coexistence of Plant Species. *Trends in Ecology and Evolution* 10:7, 263-264.
- Heller, W.D., Lindenbergh, H., Nuske, M. & Schriever, K.-H. 1978. Stochastische Systeme. De Gruyter, Berlin.
- Huston, M.A. 1994. Biological Diversity. The Coexistence of Species in Changing Landscapes. University Press, Cambridge.
- Kareiva, P. & Wennergren, U. 1995. Connecting Landscape Patterns to Ecosystem and Population Processes. *Nature* 373, 299-302.
- Kauffman, S.A. 1984. Emergent Properties in Random Complex Automata. In: Farmer, D., Toffoli, T. & Wolfram, S. (Eds.): Cellular Automata. Proceedings of an Interdisciplinary Workshop, Los Alamos, New Mexico. North-Holland Physics Publishing, Amsterdam, 145-156.
- Kummer, G., Jeltsch, F., Brandl, R. & Grimm, V. 1994. Kopplung von Prozessen auf lokaler und regionaler Skala bei der Tollwutausbreitung: Ergebnisse eines neuen Modellansatzes. *Verhandlungen der Gesellschaft für Ökologie* 23, 355-364.
- Kreeb, K.H. 1983. Vegetationskunde. Ulmer, Stuttgart.
- Marsula, R. & Ratz, A. 1994. Einfluß von Feuer auf die Populationsdynamik von serotinen Pflanzen - ein Modell. *Verhandlungen der Gesellschaft für Ökologie* 23, 365-372.
- Molofsky, J. 1994. Population Dynamics and Pattern Formation in Theoretical Populations. *Ecology* 75:1, 30-39.
- Phipps, M.J. 1992. From Local to Global: The Lesson of Cellular Automata. In: DeAngelis, D.L. & Gross, L.J. (Eds.): Individual-based Models and Approaches in Ecology. Populations, Communities and Ecosystems. Chapman & Hall, New York, 165-187.
- Ratz, A. 1994. Modellierung feuererzeugter räumlicher Strukturen in borealen Wäldern. *Verhandlungen der Gesellschaft für Ökologie* 23, 373-381.
- Silvertown, J., Holtier, S., Johnson, J. & Dale, P. 1992. Cellular Automaton Models of Interspecific Competition for Space - The Effect of Pattern on Process. *Journal of Ecology* 80, 527-534.
- Wiegand, T., Milton, S.J. & Wissel, C. 1994. Ein räumliches Simulationsmodell für eine Pflanzengemeinschaft in der südlichen Karoo, Südafrika. *Verhandlungen der Gesellschaft für Ökologie* 23, 407-416.
- Winkler, E., Klotz, S. & Wissel, C. 1994. Mathematische Modellierung der Phytozönosedynamik eines Trockenrasens. *Verhandlungen der Gesellschaft für Ökologie* 23, 451-458.
- Wolfram, S. 1983. Statistical Mechanics of Cellular Automata. *Reviews of Modern Physics* 55:3, 601-644.
- Wolfram, S. 1984. Universality and Complexity in Cellular Automata. *Physica D* 10, 1-35.

# Inconsistency of Line Convolution as a Quality Factor in the Maintenance of Spatial Databases

Michael J.C. Weir<sup>1</sup>

---

**Abstract.** — The interpretation of stand boundaries on aerial photographs has long been, and remains, an important technique of acquiring spatial data to support the management of forests and forest land. As forest management agencies move from the initial establishment of spatial databases to the task of database updating, the maintenance of data quality becomes an important issue. With time also, the task of database base updating will pass to new generations of photointerpreters. In addition to disagreements in the actual interpretation, photointerpreters differ in the level of detail with which they delineate stand boundaries and other features. Preliminary research has been undertaken to assess the extent to which photointerpretation for database updating is influenced by the convolution of earlier interpreted boundaries.

The convolution of stand boundaries independently delineated by three experienced interpreters on orthophotographs of a forest area in Austria was measured and compared. Delineations were then made for three separate areas, one by each interpreter. These initial interpretations of orthophotographs from 1974 were then updated by the other two interpreters using orthophotographs from 1984. Measurement of the convolution of the initial and updated stand boundaries suggests that individual interpreters tend to retain a characteristic approach to delineation. Consequently, it may be expected that subsequent updates by new generations of photointerpreters may result in databases containing forest stand boundaries delineated with inconsistent levels of convolution.

---

<sup>1</sup> Lecturer, International Institute for Aerospace Survey and Earth Sciences (ITC), Enschede, The Netherlands.

# Lead-Based Paint Survey Data: Spatial Variation

Tze-San Lee<sup>1</sup>

---

**Abstract.**—To identify and assess the lead-based paint (LBP), Galson Corporation was retained by the Air Force to conduct the LBP survey at twenty-four Air Combat Command Bases. Prior to starting the survey for each base, Galson's LBP Survey Manager, in consultation with Base Civil Engineering and Military Housing, reviewed information on construction and painting history for base housing. Based on this review, housing types were placed into different homogeneous exposure groups (HEGs). From each HEG, the building and the location of painted components were randomly selected to be tested. From twenty-four completed LBP surveys, four are randomly selected for an analysis. Our study shows that due to the inherent spatial variation in the LBP survey data there is a large base-to-base variability. Furthermore, most housing groups fail to satisfy the HEG assumption. By constructing choropleth maps, homogeneous subgroups within some nonhomogeneous housing groups are then identified. Choropleth maps for the housing groups at Ellsworth Air Force Base are displayed as an illustration.

---

---

<sup>1</sup>*Associate Professor, Dept. of Mathematics, Western Illinois Univ., Macomb, Ill., 61455.*

# Uncertainty of Spatially-averaged Rainfall Estimates from Rain Gauges

Jeffrey R. McCollum<sup>1</sup> and Witold F. Krajewski<sup>2</sup>

---

**Abstract.**—This study is conducted to aid the Global Precipitation Climatology Project (GPCP), which was initiated by the World Meteorological Organization. The objective of the GPCP is the production of global rainfall estimates by combining different sources of information such as satellite remote sensing, weather forecast models, and rain gauges. It is important to understand the accuracy of the rain gauge estimates because these are considered the most reliable estimates for regions with sufficient numbers of gauges. Thus, these estimates have a significant influence on the combined global rainfall products. Methods to model the error resulting from using the arithmetic mean of all the rain gauge values in a domain to estimate the spatially-averaged rainfall accumulation over the domain have been developed previously. Two of these methods are investigated in this study- an empirical equation derived from rain gauge data and an analytical equation derived from statistical concepts. Simulations are performed to determine the accuracy of these two equations in estimating the error of the rain gauge mean. A stochastic rainfall model is used to simulate rainfall over large (256 x 256 km<sup>2</sup>) domains.

---

---

<sup>1</sup> Iowa Institute of Hydraulic Research, University of Iowa, Iowa City, IA.

<sup>2</sup> Iowa Institute of Hydraulic Research, University of Iowa, Iowa City, IA.

# Defining Urban Settlements : Certainties Regarding Uncertainties

Mahavir<sup>1</sup> and Sahar Al-Amir<sup>1</sup>

---

Proper definition of urban settlements has been a consistent problem with city planners, administrators and census officers alike. The problem increases manifold, specially in developing countries, where already complex settlements are frequently appended with unplanned and unexpected developments in the fringe areas. This creates a 'data poor' situation.

Data obtained through aerial photographs and satellite images provide a better understanding of this complexity, and at a pace compatible to the rate of change in complexity. Uncertainties do remain in interpreting such data depending upon the scale and/or resolution. However, most uncertainties can be traced back to definition of such settlements, and a long waiting period (say, 10 years for a census) for a settlement to be tested against the definition. With the conventional definitions of urban settlements, uncertainties will remain even with a 100 percent ground survey.

Most accepted definitions of urban settlements are generally based on population size and/or function. There are conflicts in definitions adopted by different authorities too. For instance, while the Census of India bases the classification on population, density and employment in non-agricultural sectors, the Governor of a state may notify a settlement as urban based on generation of revenue also. Delhi (in India) is both a state (province) and a city. Where does one draw the line? Similarly, Damascus (in Syria) is both a province and a city. Should the unplanned settlements appended to the main city be subjected to the development control (promotion) policies of the city of Damascus, or the province of Damascus, or both, or none?

The technology of remote sensing offers a tool enabling administrators in providing a better understanding of this complexity in a way that can be neutrally executed by technicians. A flexible approach to definitions, i.e., the concept of 'Continuously Built-up Area', within a set of rules for operation, together with appropriate resolution and interpretation techniques help to minimize such uncertainties. While low resolution data (e.g., 72.5m of IRS-1B) help delineating, the high resolution data (e.g., 10m of SPOT, 6m of IRS-1C, and aerial photographs) support in the operationalization of this concept.

---

<sup>1</sup> *Division of Urban Survey and Human Settlement Analysis, International Institute for Aerospace Survey and Earth Sciences, Enschede, The Netherlands.*

# Spatial Dependence of Tree Biomass in an Old-Growth Forest

Franco Biondi<sup>1</sup> and Donald E. Myers<sup>2</sup>

---

**Abstract.**—Geostatistical models were fitted to aboveground tree biomass and decadal biomass increment of an old-growth, naturally-seeded stand of southwestern ponderosa pine (*Pinus ponderosa* Dougl. ex Laws. var. *scopulorum*) whose density increased over the twentieth century. Our objective was to determine the effects of crowding on spatial dependence of tree biomass, especially compared to our previous results concerning tree size and basal area increment (Biondi *et al.* 1994). Despite increased crowding, tree biomass maintained the same type of spatial dependence from 1920 to 1990. An isotropic Gaussian variogram was the model of choice to represent spatial dependence at all times. Tree biomass was spatially autocorrelated over distances no greater than 30 m, a measure of average patch diameter in this forest ecosystem. Since patch diameter remained constant through time, tree density increased by increasing the number of pine groups, not their horizontal dimension. Spatial dependence of tree biomass was greater than that of biomass increment. Increased crowding over time altered spatial dependence of biomass increment: even though the isotropic Gaussian variogram was still the model of choice, its range doubled from 1920-30 to 1980-90, while the amount of spatial dependence was halved. The latter finding indicates that decadal-scale accumulation rates over the twentieth century became less dependent on mutual tree position. These results confirmed our previous findings based on tree size and basal area increment. Simulation models and area estimates of stand dynamics in old-growth forests may be improved by including geostatistical components to summarize ecological spatial dependence.

---

## REFERENCES

- Biondi, F., D.E. Myers and C.C. Avery 1994. Geostatistically modeling stem size and increment in an old-growth forest. *Canadian Journal of Forest Research* 24: 1354-1368.

---

<sup>1</sup> *Dendrochronologist, Scripps Institution of Oceanography, University of California-San Diego, La Jolla, CA 92093-0215, USA*

<sup>2</sup> *Professor, Department of Mathematics, University of Arizona, Tucson, AZ 85721, USA*

# Sensibility and Uncertainty Analyses of an Expert System to Determine Stand Treatments

Otto ECKMUELLNER<sup>1</sup> and Martin MOSER<sup>2</sup>

---

**Abstract.-** After introducing the case study Schlaegl including possible stand treatments the sense and objective of sensibility analyses for determining a practical working program is explained.

The success of developed strategies depends on a clear formulated goal and the accuracy of the used input variables. For practical purposes the local distribution of treatments is important.

Local sensibility analyses can be used either for determining the consequences of decisions or for a ranking in the urgencies. Consequently treatment areas with similar strategies can be selected and the financial and technical need for particular areas is predictable. Furthermore these sensibility analyses provide information concerning the necessity and the level of success about sceduled strategies.

Finally three examples of possible decision criterias are presented. The first example shows possible problems for different kinds of fertilisation and the related pH-value as a used variable for selecting stand treatments. Then that the sensibility of interactions between humus thickness and soil preparation is not appropriate for general recommandations because of the grat impact of unpredictable local influences to the result. The final example - Stand Density versus thinning treatment shows a stabile behavior and is therefor suitable for determining the urgency of thinning treatments.

---

<sup>1</sup> and <sup>2</sup> Biometrician, Institute of Forest Growth, University of Bodenkultur Vienna Austria

# Stratified Two Stage Sampling (Self-Weighted) for Assessment of Village Forest Resources in Bangladesh

S. S. Islam

---

**Abstract.**—In 1994, a study of testing Stratified Two Stage Sampling (Self-Weighted) was done in village forest resources of 14 rural thanas of Chittagong district in Bangladesh. A target population of 900 rural households was formed through a preliminary survey. A field sheet was developed. Data were collected as per the field sheet. The seven parameters such as total homestead area (ha), total area (ha) of compact tree garden, total volume (m<sup>3</sup>) of trees, total volume (m<sup>3</sup>) of compact trees, total number of bamboos, volume (m<sup>3</sup>) of compact trees per hectare and number of households with bamboo plantation were calculated through complete enumeration of the target population. They were also estimated on the basis of 250 sample households through the proposed sampling design. The estimates were compared with the population values on the basis of percentage standard error, percentage accuracy, percentage bias and 95% confidence interval. Stratified Two Stage Sampling (Self-Weighted) was found to be a suitable inventory technique and was recommended to apply in the field for assessing village forest resources in Bangladesh.

---

# Modelling Animal Movements within a G.I.S. Framework

Joachim J., Janeau G. and Spitz F.\*

---

**Abstract.-** Connecting standardised animal radio-locations with a geo-referenced environmental database may point out various problems. Some aspects of the methodology will depend on specific scientific goals, but the main features of the methodology will concern most objectives. A Geographic Information System (GIS) is built up for the study area, using extant geographical data from, for instance, classified satellite imagery, topographic, tourist, ecological and vegetation maps, administrative and technical inventories, fauna, flora and soil surveys, in fact any available geo-referenced environmental database. Animal locations and segments between valid locations are vector data. They define a vector coverage through which all other data have to be taken into account. Raster and vector data in the GIS, as well as data in the animal vector coverage are adjusted in both scale and resolution. How to choose the best scale and resolution for some purpose ? The mere identity of the studied species may not be sufficient because spatial needs are often different within a same species for breeding, foraging, seeking or settling on a territory etc.... Adaptation of choices to special purposes are shown in a few examples for the Wild Boar *Sus scrofa* and the Mouflon *Ovis gmelini* in southern France.

---

---

\* INRA, Institut de Recherches sur les Grands Mammifères, BP 27, 31326  
CASTANET- TOLOSAN Cedex - France.

# Evaluation of Forest Condition Assessment Data

F.Mitterböck \*, O.Eckmüllner \*\*

---

A central part of forest condition monitoring in Austria and Europe is the visual assessment of the density of tree crowns. By no means an objective parameter, its comparability over several years highly depends on trained, 'calibrated' staff, which is only available for short terms to contribute to special research programs beside the national forest condition monitoring program. In a case study that develops concepts and means for the restoration of spruce-forest ecosystems in the Austrian Alps, distribution patterns and timely changes of crown densities provide basic data concerning the 'where, when and what' of specific restoration and management measures. It is crucial for the whole planning process to be familiar with the reliability of these data, because slight shifts in crown density can result in big differences of area when certain measures are to be applied. On an area of about 2.000 hectares, crown density on 850 sample points was assessed by two different student-teams in 1985 and 1991. This area-covering assessment was paralleled by an annual monitoring of crown density on 16 sample plots by a trained and calibrated member of the national crown assessment team. With this calibration series on hand it was possible to check the quality of the two area-covering assessments, sort out the real changes in crown density between 1985 and 1991 and to show how a specific felling policy, natural mortality, and the subjective bias of the two different assessment-teams contributes to density dynamics.

---

\* GIS-Specialist, University for Natural Resources, Vienna, Austria,  
Institute for Forest Entomology, Forest Pathology and Forest Protection

\*\* Biometrician, University for Natural Resources, Vienna, Austria, Institute for Forest Growth Research

# Application of Remote Sensing Technology in Assessing Village Wood Resources in Bangladesh

Jamil Ahmed Chowdhury<sup>1</sup>

---

**Abstract.**—Recent advances in the quality and analytical technology of satellite imagery have made it possible to achieve high precision in forest resource inventories. Increasing ground resolution in particulars, is of special interest in the study of small forest areas, and in cutting down the need for costly ground measurements. This paper presents the state-of-the-art in using remote sensing technology in assessing wood resources in Bangladesh. It presents some results of forest inventory using SPOT imagery, and discusses the need to develop the remote sensing technology towards the capability to assess wood resources in village conditions where mixed cropping is practised. Practical limitations are pointed out.

---

## INTRODUCTION

Recent developments and strategies in the forestry sector of Bangladesh are targeted self sufficiency in wood resources through social/ community forestry practice. This represents a partial shift from the traditional system of maintaining large, spatially separate and identifiable forests to a "balance" position where the bulk of the basic wood demand will be met by wood from village wood resources. The task of the resource assessment expert now has to shift towards estimating wood resources in scattered villages and homesteads.

### Preliminary Sampling Design

Each district is stratified by crop density/crop combinations and within each stratum, a double sampling procedure is applied, the roles of remote sensing data in this scenario are :

- to produce accurate maps of crop areas.
- to aid in crop stratification
- to identify the time/season suitable for crop measurement.

In order to reduce sampling costs, it is useful to consider a two or three-stage design whereby the areas which are to be stratified are themselves secondary or tertiary sampling units respectively.

## CONCLUSION

Establishment of a system of plots comparable to the permanent sample plots in villages might yield good results if used in combination with remote sensing products.

---

*Divisional Officer, Bangladesh Forest Research Institute, Chittagong, Bangladesh*

# Quality Evaluation Services on the Internet

Anders Östman<sup>1</sup>

---

**Abstract.** - Current methods for quality evaluation are often suited for data producers and not for data users. Depending on type of data and production methods used, data producers may utilize different quality evaluation techniques such as product testing, prototype testing, error propagation studies or expert evaluations. An ordinary data user is however usually producing results without having access to proper quality evaluation procedures.

A research program have been established at the University of Luleå, Sweden, aiming at developing quality evaluation procedures for casual GIS users. The approach taken here is to design and implement distributed and harmonized quality services in various environments. These services are then used as building blocks in the design of quality evaluation procedures for casual users. In this paper, two initial studies within this program are presented.

Error propagation studies and Monte Carlo simulators are important tools within quality systems for spatial data. Due to the lack of proper quality specifications of existing data sets, ordinary users are seldom using proper quality evaluation procedures. A Monte Carlo simulator for elevation data has therefore been designed and made available on the internet. The communication with clients is currently based on email and the transfer ascii files.

The MC simulator is based on the result of a large international DTM test conducted by the ISPRS. It is primarily aimed for photogrammetrically sampled digital elevation models. The client sends an email to the server, where keywords such as flying height, point density, size of grid, number of simulations and email return adress are specified. The server then generates and returns sets of error grids to the specified email adress. The values in the error grids are spatially correlated according to the results of the ISPRS DTM test and specified parameters.

Another quality service, which currently is being designed, is test procedures for quality evaluation of commercial DTM software. Test data from the ISPRS DTM test is here made public together with quality evaluation procedures of interpolated terrain surfaces.

Further information concerning this research program and more detailed information concerning the current quality services is available at the WWW-site [http://www.luth.se/depts/sb/fo\\_git/](http://www.luth.se/depts/sb/fo_git/).

---

<sup>1</sup> *Professor, Geographic Information Technology, Dept of Environmental Planning and Design, University of Luleå, S-971 87 Luleå, SWEDEN. Email: Anders.Ostman@sb.luth.se*

# Assessing Attainment of Standard by Two Stage Sampling

**Debashis Kushary**  
**Rutgers University - Camden**

---

**Abstract:** Consider assessing a process whether it complies with the regulation or not. Often it is done in two stages as described in EPA (1986) guide SW-846 (pp nine-14 to nine-17). In the first stage, sample mean of the measurements and then the upper (lower) confidence bound are computed and compared against  $L$ , the allowed maximum (minimum). If  $L$  is not in between the mean and the confidence limit, then appropriate decision is made by looking at which side  $L$  falls. If  $L$  is in between them, then a second set of samples is taken and confidence limit is recomputed using all the samples and again compared against  $L$ . If  $100(1-\alpha)\%$  confidence is desired in the procedure then a different value of  $\alpha$  should be used to maintain the over all error level at  $\alpha$  because the first sample is biased. Atwood and Bryan (Environmetrics, 1995) provided tables and curves for practitioners for normally distributed data. This paper deals with the nonparametric solution to this problem using bootstrap methodology.

---

# Site Positioning Accuracies Using GPS; Y Code Receivers and Real-time Differential Options

Karl E. Brown<sup>1</sup>

---

**Abstract.**—A majority of the current nine DOI Bureaus utilize the Department of Defense created Global Positioning System in conducting their official activities. Accuracies span a continuum from multi- to sub-meter positioning. As non-military users, civilian bureaus have been limited to the use of the Standard Positioning Service (SPS) with the effects of Selective Availability limiting [uncorrected] horizontal accuracy to approximately 60-100 meters for real-time point positioning. Current differential correction practices can correct field points to 1-3 meters accuracy, either after the fact, or in real-time over low power radio links, or by subscription to a correction service. Comparatively, the Precise Positioning Service (PPS) delivers 16 meter horizontal accuracy anywhere in the world, with typical North American positioning experience at 10 meters or better. The authority to utilize the PPS was granted to DOI by the Dept. of Defense (DoD) on Aug. 26, 1994. Unfortunately, although commercial real-time broadcast services can allow a user to obtain 1-3 meters positioning accuracy, a current significant lapse in service into many wildland areas does not fully support the needs of the DOI Bureaus. Essentially, these services are based on urban demand and urban subscription income, with little evidence of future planning to support users on the remote lands administered or studied by DOI. *Since Natural Resources and Environmental study sites may occur in either wildland or urban environments, both autonomous PPS, and SPS receivers with real-time differential can meet the positioning needs of DOI users needing to map, inventory, or obtain georeferenced data on these sites.* Current studies for both the US Coast Guard and the Federal Aviation Administration have proposed real-time GPS augmentation systems. Again, the proposed systems are not sufficient to reach the *wildland* DOI users until full augmentation by the FAA occurs in late 1998. That full system ASSUMES approval of wide area correction signal distribution. The DoD maintains that such a system might violate National Security issues, and currently does not support the distribution of *unencrypted differential correctors over a real-time system.* The DoD has recognized a positioning need in the civilian agencies for 16 meter accuracy, and has made the PPS and the military PLGR receiver available to Federal civilian users. Other PPS capable receivers are available in addition to the PLGR; however, they are more expensive. In order to use the PPS, an encryption key is utilized, and the receivers must be keyed annually. The National Security Agency (NSA) manages the keying materials. The keyed material is classified as Secret [crypto], and a chain of custody with appropriate clearances is required. Typically, NSA wants to limit the number of users of keyed material, so creating a DOI cooperative site in 1995 would allow the DOI to concentrate it's access for keyed material to a secure site while still delivering the needed capabilities to the field. A full range of accuracies are available to DOI users, depending on varying costs, hardware and software installations, and training/security requirements.

---

<sup>1</sup>Nationwide GPS Coordinator, USDI National Biological Service, Denver, CO

# Statistical Models of Landscape Pattern and the Effects of Coarse Resolution of Satellite Imagery on Estimation of Area

Christine A. Hlavka<sup>1</sup>

---

**Abstract.**—Analysis of classified satellite imagery was conducted to characterize errors in estimates of area based on coarse resolution satellite imagery which are due to distortions in sizes of small fragments, and to explore the feasibility of correcting for these errors using a statistical modeling approach. Sizes of bodies of open water on ERS-1 SAR and fire scars on Landsat MSS imagery were measured. Statistical analysis of the smaller scars and ponds as observed with this imagery of relatively fine resolution demonstrated that the distribution of the sizes could be modeled by either of two types of statistical distributions - a power distribution related to fractal processes or a simple exponential distribution. Comparison of the distribution of small burn scars as observed with Landsat to the distribution observed with AVHRR showed distortions due to the coarse spatial resolution of AVHRR caused a net overestimation of burn area. This bias was primarily caused by detection in 2 or 3 AVHRR pixels of burns whose true size was on the order of an AVHRR pixel.

---

---

<sup>1</sup>NASA/Ames Research Center.

# Author Index

Achard, F. ....	366	Fatale, L.A. ....	301	Kuo, W.L. ....	65
Ackeret, J.R. ....	301	Ferri, F. ....	623	Kushary, D. ....	724
Adler, P. ....	267	Fisher, P.F. ....	272, 585	Lal, J.B. ....	610
Al-Amir, S. ....	716	Fitzgerald, R.W. ....	317	Lane, L.J. ....	147
Anderson, M.C. ....	561	Forier, F. ....	225	Lee, T-S. ....	714
Antle, A.N. ....	593	Fosnight, E.A. ....	425	Lewis, A. ....	317
Avena, G. ....	623	Fowler, G.W. ....	425	Lewis, A. ....	577
Baltzer, H. ....	695, 703	Gallego, F.J. ....	509	Limp, W.F. ....	517
Besag, J. ....	493	Garner, M.E. ....	517	Liu, Z. ....	129
Biondi, F. ....	717	Gertner, G.Z. ....	639	Lodin, M. ....	327
Bogardi, I. ....	49	Goderya, F. ....	49	Lowell, K. ....	235
Bolstad, P.V. ....	359	Goodchild, M.F. ....	217, 568	Mackey, B.G. ....	291
Borgman, L.F. ....	155	Gotway, C.A. ....	30	Mahavir ....	716
Boucneau, G. ....	251	Gregoire, T.G. ....	667	Malingreau, J-P. ....	366
Bowden, D.C. ....	694	Griffith, D.A. ....	147	Marshall, P.L. ....	593
Bowser, G. ....	676	Gross, C-P. ....	267	May, J.A. ....	4
Brand, G.J. ....	173	Grunig, A. ....	441	McCollum, J.R. ....	715
Braun, P. ....	695,703	Guan, B.T. ....	639	McConville, D. ....	351
Brion, P. ....	617	Gumpertz, M. ....	401	McCulloch, C.E. ....	65
Brockhaus, J. ....	401	Gutter, P. ....	115	McKenney, D.W. ....	291
Brown, A. ....	189	Hall, J. ....	199	McRoberts, R.E. ....	659
Brown, K.E. ....	725	Hallberg, G. ....	129	Messmore, J.A. ....	301
Burgan, R.E. ....	409	Hasenauer, H. ....	667	Milliken, J.A. ....	541
Cahoon, D.R. Jr. ....	383	Hazelhoff, L. ....	601	Mitterbock, F. ....	721
Campbell, J. ....	508	Heemink, A.W. ....	181	Moffett, J.L. ....	493
Canters, F. ....	225	Hershey, R.R. ....	73	Mohler, C. ....	65
Casey, K. ....	343	Heuvelink, G.B.M. ....	243	Moisen, G.G. ....	459
Catanzaro, D.G. ....	517	Hlavka, C.A. ....	726	Monserud, R.A. ....	667
Chen, X. ....	545	Hock, J. ....	217	Moody, A. ....	83
Chowdhury, J.A. ....	722	Hoefsloot, F. ....	601	Moser, M. ....	718
Cieszewski, C.J. ....	649	Hoeting, J. ....	694	Murillo, M.L. ....	281
Clerke, W. ....	508	Hofman, G. ....	251	Myers, D.E. ....	23, 717
Coburn, T.C. ....	4	Holdaway, M.R. ....	173	Myers, W.L. ....	561
Collins, J.B. ....	121	Howarth, P.J. ....	91	Nasu, M. ....	545
Combs, R. Jr. ....	359	Hughes, J.P. ....	115	Nychka, D. ....	137
Conroy, M.J. ....	685	Huisman, J.A. ....	243	O'Neill, R.V. ....	560
Cressie, N.A. ....	1	Hunter, G.J. ....	217, 281	Ohlen, D.O. ....	409
Crocker, J.E. ....	685	Hutchinson, M.F. ....	291, 577	Ostman, A. ....	723
Cutler, D.R. ....	459	Islam, S.S. ....	719	Owens, T. ....	351
Czaplewski, R.L. ....	419,409	Jacquez, G.M. ....	259	Parysow, P. ....	639
Czaplewski, R.L. ....	467,508	Janeau, G. ....	720	Patil, G.P. ....	569
Dahab, M. ....	49	Jeanjean, H. ....	366, 375	Pelkie, C. ....	65
Daosheng, D. ....	189	Jiang, H. ....	527	Pelkki, M.H. ....	336
De Boissezon, H. ....	375	Joachim, J. ....	720	Phillips, D.L. ....	649
Dees, M. ....	391	Jones, K.B. ....	560	Pontius, J.S. ....	434
DeGloria, S. ....	65	Kalkhan, M.A. ....	467	Qian, S.S. ....	164
Doihara, T. ....	545	Kangas, A. ....	631	Quintanilha, J.A. ....	501
Doyle, P.G. ....	147	Kern, J.W. ....	155	Ranneby, B. ....	451
Drummond, J. ....	189	Khorram, S. ....	401	Reich, R.M. ....	467
Dzur, R.S. ....	517	Kleinn, C. ....	391	Ricotta, C. ....	623
Eastman, J.R. ....	527	Koebbe, J. ....	107	Riitters, K.H. ....	560
Eckmullner, O. ....	718, 721	Kohl, M. ....	441	Ritman, K.T. ....	317
Ediriwickrema, J. ....	401	Kohler, W. ....	695, 703	Robey, M. ....	217
Edwards, T.C. Jr. ....	459	Kohley, T. ....	693	Rodrigues, M. ....	501
Evans, B.M. ....	561	Korhonen, K.T. ....	631	Rogowski, A.S. ....	57
Evans, D.L. ....	419	Krajewski, W.F. ....	715	Royle, J.A. ....	137
Fahringer, J. ....	508	Kumar, P. ....	99	Rutherford, B.M. ....	30,39

Treitz, P.M. ....	91	Voyadgis, D.E. ....	208	Williams, M.T. ....	477
Tsai, F. ....	65	Wade, T.G. ....	560	Woldt, W. ....	49
Turner, D.P. ....	649	Wadsworth, R. ....	199	Woodcock, C.E. .121, 535, 541	
Van Caster, M. ....	694	Waller, L.A. ....	259	Yarus, J.M. ....	4
van Elzakker, C. ....	189	Weinstein, D.A. ....	65	Zhang, X.F. ....	181
van Hees, W.W.S. ....	553	Weir, M.J.C. ....	309,713	Zhu, Z. ....	409
Van Meirvenne, M. ....	251	Wickham, J.D. ....	560	Zimmerman, D. ....	129

## **Symposium Steering Committee**

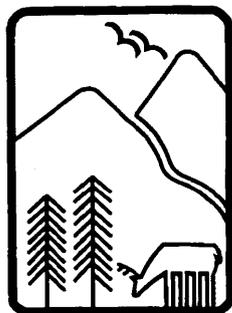
H. Todd Mowrer, Chair  
Raymond L. Czaplewski, Co-Chair  
James L. Smith, Past Chair  
Michael J.C. Weir, International Coordinator

## **Science Advisory Board**

Joseph K. Berry, Berry and Associates  
Harold E. Burkhart, Virginia Polytechnic Institute and State University  
Russell G. Congalton, University of New Hampshire  
Noel A.C. Cressie, Iowa State University  
George Z. Gertner, University of Illinois  
Richard O. Gilbert, Batelle Laboratories  
Michael F. Goodchild, University of California - Santa Barbara  
Gerard B.M. Heuvelink, University of Amsterdam  
Christopher Kleinn, Universität Freiburg  
Barbara Koch, Albert-Ludwigs-Universität  
Michael Köhl, Swiss Federal Institute for Forest, Snow, and Landscape Research  
Robert O. Kuehl, University of Arizona  
Kim Lowell, Université Laval  
H. Gyde Lund, USDA Forest Service Inventory and Analysis  
Ann Maclean, Michigan Technological University  
Anthony Olsen, US Environmental Protection Agency  
G.P. Patil, Pennsylvania State University  
Denice Shaw, US Environmental Protection Agency  
J.P. Skovsgaard, Danish Forest and Landscape Research Institute  
Steve Stehman, University of New York - Syracuse  
Mike Story, National Biological Service

The United States Department of Agriculture (USDA) prohibits discrimination in its programs on the basis of race, color, national origin, sex, religion, age, disability, political beliefs and marital or familial status. (Not all prohibited bases apply to all programs.) Persons with disabilities who require alternative means for communication of program information (braille, large print, audiotape, etc.) should contact the USDA Office of Communications at (202) 720-2791.

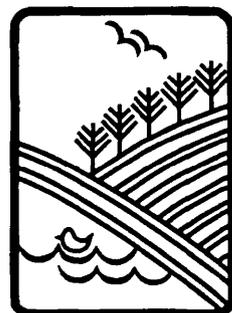
To file a complaint, write the Secretary of Agriculture, U.S. Department of Agriculture, Washington, D.C. 20250, or call (202) 720-7327 (voice) or (202) 720-1127 (TDD). USDA is an equal employment opportunity employer.



Rocky  
Mountains



Southwest



Great  
Plains

U.S. Department of Agriculture  
Forest Service

## Rocky Mountain Forest and Range Experiment Station

The Rocky Mountain Station is one of eight regional experiment stations, plus the Forest Products Laboratory and the Washington Office Staff, that make up the Forest Service research organization.

### RESEARCH FOCUS

Research programs at the Rocky Mountain Station are coordinated with area universities and with other institutions. Many studies are conducted on a cooperative basis to accelerate solutions to problems involving range, water, wildlife and fish habitat, human and community development, timber, recreation, protection, and multiresource evaluation.

### RESEARCH LOCATIONS

Research Work Units of the Rocky Mountain Station are operated in cooperation with universities in the following cities:

Albuquerque, New Mexico  
Flagstaff, Arizona  
Fort Collins, Colorado\*  
Laramie, Wyoming  
Lincoln, Nebraska  
Rapid City, South Dakota

\*Station Headquarters: 240 W. Prospect Rd., Fort Collins, CO 80526