

Environmental and Ecological Statistics **10**, 301–308, 2003

Introduction to special issue on map accuracy

STEPHEN V. STEHMAN¹ and RAYMOND L. CZAPLEWSKI²

¹*SUNY ESF, 320 Bray Hall, Syracuse, NY 13210*

²*U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, 2150 Center Avenue, Bldg. A, Suite 350, Fort Collins, CO 80526*

1. Background

With the advent of satellite remote sensing and computing technology, mapping land cover over extensive regions of the earth has become practical and cost effective. For example, land-cover maps have been produced covering pan-Europe (Mucher *et al.*, 2000), Great Britain (Fuller *et al.*, 1994), Canada (Cihlar *et al.*, 1999), Mexico (Mas *et al.*, 2002) the United States (Vogelmann *et al.*, 2001), and the globe (Belward *et al.*, 1999). Franklin and Wulder (2002) assemble a diverse array of other examples of large-area, land-cover maps. Land-cover maps are typically an intermediate product, used, for example, as input into various hydrological and carbon cycling models (e.g., Riley *et al.*, 1997) or habitat suitability models that quantify relationships between land cover and wildlife abundance. Another common application is to quantify and map land-cover change (Lunetta and Elvidge, 1998; Donoghue, 2002), focusing on forest change (e.g., Hayes and Sader, 2001), urban development (Clarke *et al.*, 1996), or wetland loss (Jensen *et al.*, 1995). Analyzing landscape pattern metrics is still another common application of land-cover information (e.g., Wickham and Norton, 1994), and relationships between landscape pattern and biological and hydrological phenomena may be investigated (e.g., Jones *et al.*, 2001; Lawler and Edwards, 2002).

These applications are predicated on the assumption that the land-cover map is sufficiently accurate to justify its intended use. Consequently, a scientifically credible assessment of the map's accuracy is critical. The traditional approach to accuracy assessment involves three primary components, the sampling design to determine which subregions (e.g., pixels or land-cover polygons) or points will be sampled, the response or measurement design to obtain the true or "reference" attribute for each sampled unit or point, and the analysis of the data obtained (Stehman and Czaplewski, 1998). The four articles contained in this special issue span diverse issues related to accuracy, yet they integrate these three major components of accuracy assessment to create a comprehensive view of the topic. Nusser and Klaas describe a sampling strategy and results for an accuracy assessment of an Iowa land-cover map. Their methodology represents an excellent example of an accuracy assessment in which a thorough, survey-sampling approach has been employed, encompassing sample size planning, implementing a probability sampling design, providing estimates incorporating non-response adjustments, and reporting standard errors. Sampling design issues are also addressed in Steele *et al.*,

but from a different perspective. They tackle the problem of how to conduct a statistically defensible accuracy assessment when cost and/or other practical constraints prevent implementation of a probability sampling design. In addition to deriving estimates of accuracy from a non-probability sample, they develop procedures for producing an accuracy map. Patil and Taillie focus exclusively on the analysis component of an accuracy assessment strategy. They introduce the idea of latent truth analysis as a means of characterizing map accuracy. Their approach offers an interesting paradigm shift from the analyzes typically conducted in which the focus is on general descriptive summary measures and chance-corrected agreement. Lastly, Arbia *et al.* address the problem of error propagation in maps based on vegetation indices derived from linear combinations of spectral information. This article focuses on potential sources or causes of error in the classified product. The analyses developed by Arbia *et al.* take into account the effects of both positional and attribute error on the resulting classified map (e.g., a map representing a vegetation index). Although these articles focus on land cover, they address problems common to categorical maps of any kind.

The four articles may be viewed within a more extended statistical context. Nusser and Klaas's work serves as an example of how to rigorously follow survey-sampling procedures in the design and analysis of a real, practical sampling problem. In addition to ideas pertaining to use of data from a non-probability sample, Steele *et al.* discuss techniques of general applicability to classification problems, assessment of classifier error rates, and visualizing spatial patterns of error. Patil and Taillie provide a novel application of latent truth analysis, accompanied by development of the requisite estimation methodology. And finally, Arbia *et al.* present work of general interest in the analysis of error propagation. These four articles have something of interest for practitioners working in subject areas outside of environmental science, and statisticians addressing related practical applications of statistical theory and methodology should find something to satisfy their curiosity.

In the remainder of this Introduction, we provide additional background on the traditional practice of accuracy assessment, and then place the articles presented in this special issue within a more detailed context of the current state of accuracy assessment practice.

2. Measurement design

The underlying premise of an accuracy assessment is a location-specific comparison of the attribute or label displayed on the map with the "true" attribute. The true land cover of a location (e.g., point or spatial unit) may be determined by ground visit, from aerial photography or videography, or higher-resolution satellite imagery. However, the "true" classification of a location can vary depending upon the choice of definitions and measurement protocols. Therefore, the phrase "ground truth" is avoided in favor of "reference" data. Although the reference land-cover labels are expected to be more accurate than the map labels, the implication that we know the true land cover is avoided by employing the term "reference data". Similarly, the term "agreement" is sometimes preferred to "accuracy" to reflect this feature of the reference data. Patil and Taillie's latent truth analysis may be seen as a way to formally recognize and account for the inaccuracies of reference data.

The measurement design is strongly linked to the land-cover classification scheme selected for the map. The classification schemes discussed in these four articles are so-called “crisp” or “hard” classifications in which each pixel is assigned to a single land-cover class. In practice, fuzzy set concepts are sometimes employed in the classification scheme of the map and/or the reference data. The map classification may be fuzzy in the sense that each mapping unit is assigned a membership value for each of the possible land-cover classes, rather than assigned to one and only one map class (Foody, 1999). Whether or not the map employs a fuzzy classification, the reference data protocol for assessing accuracy is sometimes based on a fuzzy classification. Gopal and Woodcock (1994) introduced a fuzzy linguistic scale rating for reference data when assessing the accuracy of a map based on a crisp land-cover classification scheme. This approach to accuracy assessment is not discussed in the special issue, but it is an important topic to which statisticians may contribute useful new insights and methods. Foody (2002) reviews methods for assessing fuzzy class maps as well as techniques for assessing a crisp classification scheme using a fuzzy-class reference data protocol. Gill *et al.* (2000) and Laba *et al.* (2002) are examples in which the latter technique has been applied in practice.

3. Sampling design

Sampling design plays a critical role in accuracy assessment. Because high quality reference data are difficult and expensive to obtain, the sampling design issues encountered in accuracy assessment are similar to those traditionally addressed by survey sampling methodology: how to sample in a cost-effective, yet statistically rigorous manner. Application of basic sampling designs such as simple random, stratified random, systematic, and cluster have been summarized in various accuracy assessment review articles (e.g., Congalton, 1991; Janssen and van der Wel, 1994; Congalton and Green, 1999; Stehman, 1999; Czaplewski, 2000; Foody, 2002). These basic designs serve well for small-area land-cover maps, but they are inadequate given the practical realities of assessing the accuracy of large-area, land-cover maps. Two design criteria are typically desired for large-area map assessments. Cost limitations often dictate that cluster sampling must be used to reduce travel costs for ground visits or to reduce aerial photography costs. At the same time, assessment objectives require stratification by land-cover class to obtain adequate precision for class-specific accuracy estimates. The accuracy assessment literature provides little guidance to accommodate these dual objectives. One of the appealing features of Nusser and Klaas's design is that it does accommodate both of these desirable design criteria. Yang *et al.* (2001) provide another example employing a similar design structure.

Denied access to sampling locations is prevalent if the reference data are obtained via ground visit. Private landowner refusal and difficult or dangerous access are common reasons for denied access. Nusser and Klaas recognize and address this practical reality up front. Their survey sampling approach includes sample size calculations that factor in non-response at the planning stage, and they account for the reality of missing data at the analysis stage. Steele *et al.* also directly confront the problem of rigorous analysis in the extreme case in which no probability sampling design is implemented.

4. Analysis

Once the reference data are in hand, the next step in accuracy assessment is analysis of these data. The traditional analysis of accuracy assessment data begins with an error matrix, sometimes also called an agreement or confusion matrix (Story and Congalton, 1986). An error matrix summarizes the correct classifications and misclassifications in a contingency table format, with the rows designating the map labels and the columns the reference labels (this is the common row and column convention, but sometimes the designations are reversed). The (i, j) cell entry of the error matrix, p_{ij} , is the proportion of area that is map class i and reference class j . These proportions are estimated from the sample data, and overall accuracy of the map is derived from the diagonal elements of the error matrix. Various conditional probabilities may also be calculated from this error matrix. In the accuracy assessment jargon, “user’s accuracy” is the conditional probability of correctly classifying a location given that it has been mapped as class i , and “producer’s accuracy” is the conditional probability of having correctly mapped a location given that it is truly class j . Typically, formulas for estimating the error matrix are provided for simple random sampling, but standard error formulas are omitted (cf. Congalton, 1991; Janssen and van der Wel, 1994). Czaplewski (1994) provides general estimation formulas, including standard errors, for a variety of designs and estimators commonly used in accuracy assessment.

Early in the development of accuracy assessment analyses, chance-corrected measures of agreement were promoted both for description of individual error matrices and for comparison of error matrices. The kappa coefficient was advocated early on (Congalton *et al.*, 1983), with more recent suggestions including tau (Ma and Redmond, 1995) and weighted kappa (Naesset, 1996). The near universal acquiescence to using kappa in the practice of accuracy assessment has to some extent stunted development of alternative, perhaps more meaningful analyses. This is unfortunate because in other subject areas, kappa has undergone more scrutiny (Uebbersax, 1987; Zwick, 1988) and is not always viewed as favorably as it is in remote sensing applications (Stehman, 1997). For example, the appropriateness of comparing kappa coefficients has been questioned because of kappa’s strong dependence on the marginal distributions of the contingency tables (Agresti *et al.*, 1995). The approach of modeling agreement (Tanner and Young, 1985) versus simply describing agreement has not gained a foothold in the accuracy assessment literature. Agresti (1989) demonstrates that the model underlying kappa is of dubious practical value. As yet, little effort has been allocated to these potentially insightful, model-based analyses of error matrices. Instead, current practice follows an obligatory reporting of kappa with little questioning of its interpretive value.

The articles in this special issue illustrate new developments in the analysis of accuracy assessment data. The novelty of the Patil and Taillie approach is that it goes beyond the typical error matrix analysis to examine the underlying structure of agreement. Their latent structure approach represents a conceptually new way to think about analysis of accuracy data and goes beyond simply reporting kappa. Nusser and Klaas conduct a traditional accuracy assessment analysis, estimating the error matrix and associated summary measures. Their analysis extends beyond those typically found in the accuracy assessment literature because they take the important step of calculating standard errors for the estimates. Although it may be surprising, many published accuracy assessments do not

include standard errors with the accuracy estimates. Nusser and Klaas take advantage of the survey sampling variance estimation procedures available in SAS. This adoption of a standard software package to estimate standard errors is one of the first examples, if not the first, such implementation in a large-scale accuracy assessment. Steele *et al.* contribute an innovative analysis via their methodology for mapping accuracy. McGwire and Fisher (2001) advocate spatial representations of accuracy, and early attempts in this direction include Steele *et al.* (1998) and Kyriakidis and Dungan (2001).

5. Summary

Current practice of sampling design and analysis in accuracy assessment all too often relies on *ad hoc* non-probability sampling designs, questionable replacement strategies to remedy denied access, and analyzes that treat data from complex designs as if they arose from a simple random sample. Congalton and Green's (1999) Chapter 8 case study typifies this approach. In current accuracy assessment practice, the intent exists to implement statistically rigorous sampling strategies, but 100% success of this intent has not been achieved. Examples moving in the right direction toward cost-effective, statistically defensible strategies include Nusser and Klaas, Edwards *et al.* (1998), Scepan (1999), and Zhu *et al.* (2000).

The four articles contained in this special issue illustrate both sound fundamental methodology as well as new innovations for analysis that will further strengthen the statistical formulation of accuracy assessment. Nusser and Klaas supply a thorough survey sampling approach to the full design and analysis of an accuracy assessment protocol. Steele *et al.* develop rigorous techniques for analyzing data from a non-probability sampling design, thus providing a statistically defensible option for dealing with this practical reality of some accuracy assessments. Patil and Taillie offer a different direction from the traditional analysis of error matrices by introducing latent truth methods to the toolbox of accuracy assessment analyses. And finally, Arbia *et al.* supply new results on the propagation of error through linear vegetation indices that provide insight into sources of error in the resulting classified maps based on these indices. The developments in accuracy assessment and related methodology found in this special issue should help practitioners to construct more statistically rigorous, practical assessments of map accuracy, and to implement analyses yielding better insights and understanding of the nature of error and error propagation of these maps. We commend the authors for their worthy contributions to this practically important endeavor.

Acknowledgments

We would like to express our sincere thanks to the referees who participated in the peer-review process. Their careful attention to detail, thoughtful comments, and professional judgment significantly contributed to this special issue. With the exception of one reviewer, all reviews were solicited and obtained with author information removed. We thank Editor-in-Chief G. P. Patil for proposing the theme of this special issue and inviting us to serve as Guest Editors.

References

- Agresti, A. (1989) An agreement model with kappa as parameter. *Statistics and Probability Letters*, **7**, 271–3.
- Agresti, A., Ghosh, A., and Binia, M. (1995) Raking kappa: Describing potential impact of marginal distributions on measures of agreement. *Biometrical Journal*, **37**, 811–20.
- Belward, A.S., Estes, J.E., and Kline, K.D. (1999) The IGBP-DIS global 1-km land-cover data set DISCover: A project overview. *Photogrammetric Engineering and Remote Sensing*, **65**, 1013–20.
- Cihlar, J., Beaubien, J., Latifovic, R., and Simard, G. (1999) *Land cover of Canada 1995 Version 1.1. Digital data set documentation*, Ottawa: Natural Resources Canada.
- Clarke, K.C., Gaydos, L., and Hoppen, S. (1996) A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay area. *Environment and Planning B*, **24**, 247–61.
- Congalton, R.G. (1991) A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, **37**, 35–46.
- Congalton, R.G. and Green, K. (1999) *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, Lewis Publishers, Boca Raton, Florida.
- 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**, 1671–8.
- Czaplewski, R.L. (1994) Variance approximations for assessments of classification accuracy, Res. Pap. RM-316, U.S. Department of Agriculture, Forest Service, Rocky Mountain Forest and Range Experiment Station, Fort Collins, CO, p. 29.
- Czaplewski, R.L. (2000) Accuracy assessments and areal estimates using two-phase stratified random sampling, cluster plots, and the multivariate composite estimator. In *Quantifying Spatial Uncertainty in Natural Resources*, H.T. Mowrer and R.G. Congalton (eds), Ann Arbor Press, Chelsea, Michigan. pp. 79–100.
- Donoghue, D.N.M. (2002) Remote sensing: environmental change. *Progress in Physical Geography*, **26**, 144–51.
- Edwards, T.C., Jr., Moisen, G.G., and Cutler, D.R. (1998) Assessing map accuracy in a remotely-sensed ecoregion-scale cover-map. *Remote Sensing of Environment*, **63**, 73–83.
- Foody, G.M. (1999) The continuum of classification fuzziness in thematic mapping. *Photogrammetric Engineering and Remote Sensing*, **65**, 443–51.
- Foody, G.M. (2002) Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, **80**, 185–201.
- Franklin, S.E. and Wulder, M.A. (2002) Remote sensing methods in medium spatial resolution satellite data land cover classification of large areas. *Progress in Physical Geography*, **26**, 173–205.
- Fuller, R.M., Groom, G.B., and Jones, A.R. (1994) The landcover map of Great Britain: an automated classification of Landsat Thematic Mapper data. *Photogrammetric Engineering and Remote Sensing*, **60**, 553–62.
- Gill, S.J., Milliken, J., Beardsley, D., and Warbington, R. (2000) Using a mensuration approach with FIA vegetation plot data to assess the accuracy of tree size and crown closure classes in a vegetation map of Northeastern California. *Remote Sensing of Environment*, **73**, 298–306.
- Gopal, S. and Woodcock, C. (1994) Theory and methods for accuracy assessment of thematic maps using fuzzy sets. *Photogrammetric Engineering and Remote Sensing*, **60**, 181–8.
- Hayes, D.J. and Sader, S.A. (2001) Comparison of change-detection techniques for monitoring tropical forest clearing and vegetation regrowth in a time series. *Photogrammetric Engineering and Remote Sensing*, **67**, 1067–75.

- Janssen, L.L.F. and van der Wel, F.J.M. (1994) Accuracy assessment of satellite derived land-cover data: A review. *Photogrammetric Engineering and Remote Sensing*, **60**, 419–26.
- Jensen, J.R., Rutchey, K., Koch, M., and Narumalani, S. (1995) Inland wetland change detection in the Everglades Water Conservation Area 2a using a time series of normalized remotely sensed data. *Photogrammetric Engineering and Remote Sensing*, **61**, 199–209.
- Jones, K.B., Neale, A.C., Nash, M.S., Van Remortel, R.D., Wickham, J.D., Riitters, K.H., and O'Neill, R.V. (2001) Predicting nutrient and sediment loadings to streams from landscape metrics: a multiple watershed study from the United States Mid-Atlantic region. *Landscape Ecology*, **16**, 301–12.
- Kyriakidis, P.C. and Dungan, J.L. (2001) A geostatistical approach for mapping thematic classification accuracy and evaluating the impact of inaccurate spatial data on ecological model predictions. *Environmental and Ecological Statistics*, **8**, 311–30.
- Laba, M., Gregory, S.K., Braden, J., Ogurcak, D., Hill, E., Fegraus, E., Fiore, J., and DeGloria, S.D. (2002) Conventional and fuzzy accuracy assessment of the New York Gap Analysis Project land cover maps. *Remote Sensing of Environment*, **81**, 443–55.
- Lawler, J.J. and Edwards, T.C. (2002) Landscape patterns as habitat predictors: building and testing models for cavity-nesting birds in the Uinta Mountains of Utah, USA. *Landscape Ecology*, **17**, 233–45.
- Lunetta, R.S. and Elvidge, C.D. (1998) *Remote Sensing Change Detection: Environmental Monitoring Methods and Applications*, Sleeping Bear Press, Inc., Chelsea, MI.
- Ma, Z. and Redmond, R.L. (1995) Tau coefficients for accuracy assessment of classification of remote sensing data. *Photogrammetric Engineering and Remote Sensing*, **61**, 435–9.
- Mas, J.-F., Velázquez, A., Palacio-Prieto, J.L., Bocco, G., Peralta, A., and Prado, J. (2002) Assessing forest resources in Mexico: Wall-to-wall land use/cover mapping. *Photogrammetric Engineering and Remote Sensing*, **68**(10), 966–8.
- McGwire, K.C. and Fisher, P. (2001) Spatially variable thematic accuracy: Beyond the confusion matrix. In *Spatial Uncertainty in Ecology: Implications for Remote Sensing and GIS Applications*, C.T. Hunsaker, M.F. Goodchild, M.A. Friedl, and T.J. Case (eds.), Springer, New York. pp. 308–29.
- Mücher, C.A., Steinnocher, K.T., Kressler, F.P., and Heunks, C. (2000) Land cover characterization and change detection for environmental monitoring of pan-Europe. *International Journal of Remote Sensing*, **21**, 1159–81.
- Naesset, E. (1996) Use of the weighted Kappa coefficient in classification error assessment of thematic maps. *International Journal of Geographic Information Systems*, **10**, 591–604.
- Riley, R.H., Phillips, D.L., Schuft, M.J., and Garcia, M.C. (1997) Resolution and error in measuring land-cover change: effects on estimating net carbon release from Mexican terrestrial ecosystems. *International Journal of Remote Sensing*, **18**, 121–37.
- Scepan, J. (1999) Thematic validation of high-resolution global land-cover data sets. *Photogrammetric Engineering and Remote Sensing*, **65**, 1051–60.
- Steele, B.M., Winne, J.C., and Redmond, R.L. (1998) Estimation and mapping of misclassification probabilities for thematic land cover maps. *Remote Sensing of Environment*, **66**, 192–202.
- Stehman, S.V. (1997) Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, **62**, 77–89.
- Stehman, S.V. (1999) Basic probability sampling designs for thematic map accuracy assessment. *International Journal of Remote Sensing*, **20**, 2423–41.
- Stehman, S.V. and Czaplewski, R.L. (1998) Design and analysis for thematic map accuracy assessment: Fundamental principles. *Remote Sensing of Environment*, **64**, 331–44.
- Story, M. and Congalton, R.G. (1986) Accuracy assessment: a user's perspective. *Photogrammetric Engineering and Remote Sensing*, **52**, 397–9.
- Tanner, R. and Young, M.A. (1985) modeling agreement among raters. *Journal of the American Statistical Association*, **80**, 175–80.

- Uebersax, J.S. (1987) Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin*, **101**, 140–6.
- Vogelmann, J.E., Howard, S.M., Yang, L., Larson, C.R., Wylie, B.K., and Van Driel, N. (2001) Completion of the 1990s National Land Cover Data set for the conterminous United States from Landsat Thematic mapper data and ancillary data sources. *Photogrammetric Engineering and Remote Sensing*, **67**, 650–62.
- Wickham, J.D. and Norton, D.J. (1994) Mapping and analyzing landscape patterns. *Landscape Ecology*, **9**, 7–23.
- Yang, L., Stehman, S.V., Smith, J.H., and Wickham, J.D. (2001) Thematic accuracy of MRLC land cover for the Eastern United States. *Remote Sensing of Environment*, **76**, 418–22.
- Zhu, Z., Yang, L., Stehman, S.V., and Czaplewski, R.L. (2000) Accuracy assessment for the U.S. Geological Survey regional land-cover mapping Program: New York and New Jersey region. *Photogrammetric Engineering and Remote Sensing*, **66**, 1425–35.
- Zwack, R. (1988) Another look at interrater agreement. *Psychological Bulletin*, **103**, 374–8.

Biographical sketches

Stephen V. Stehman is an Associate Professor of Biometry in the Department of Forest and Natural Resources Management at SUNY ESF. Along with Ray Czaplewski, he was a member of a team of U.S. Environmental Protection Agency and U.S. Geological Survey scientists who conducted the accuracy assessment of the 1992 National Land-Cover Data (NLCD), a land-cover map of the conterminous United States. His current interest is applying statistical methods to problems of map accuracy assessment.

Raymond L. Czaplewski is Project Leader for the Forest Inventory and Monitoring Environmetrics Research Unit at USDA Forest Service's Rocky Mountain Research Station in Fort Collins, Colorado, USA. This Unit uses mathematical statistics to address national issues related to the Forest Service's Forest Inventory and Analysis (FIA) program. FIA is the "Census of the Nation's forests." In addition to statistical products, FIA and its cooperators build geospatial datasets for the nation's forests through combination of FIA field data with remotely sensed data and other geospatial information. FIA requires accuracy assessments for all of its geospatial products.