Accuracy Assessment with Complex Sampling Designs

Raymond L. Czaplewski
United States Forest Service
Rocky Mountain Research Station
Fort Collins, Colorado USA
rczaplewski@fs.fed.us

***Abstract*— A reliable accuracy assessment of remotely sensed geospatial data requires a sufficiently large probability sample of expensive reference data. Complex sampling designs reduce cost or increase precision, especially with regional, continental and global projects. The General Restriction (GR) Estimator and the Recursive Restriction (RR) Estimator separate a complex sample survey into simple statistical components, each of which is sequentially combined into the final estimate. GR and RR produce a design-consistent Empirical Best Linear Unbiased Estimator (EBLUE) for any sample survey design, regardless of its complexity.**

Keywords: Kalman filter; error matrix; GIS; geospatial database; MODIS; Landsat; LiDAR; photo-interpretation

I. INTRODUCTION

An accuracy assessment provides essential metadata for geospatial databases that are produced with remotely sensed data, especially at the regional, continental and global scales. Strong inference about database accuracy requires a valid probability sampling design, a precisely defined sample-population, sufficient sample size, relevant reference data, and a design-consistent estimator (Czaplewski 2003).

For example, the correlation coefficient is a common accuracy assessment statistic for continuous variables (*e.g.*, biomass). A simple random sample of point-plots is the simplest probability sampling design. Each point-plot has two known values: the reference datum (*e.g.*, a single field measurement of biomass); and the predictor datum (*e.g.,* a single prediction of biomass with remotely sensed data). The correlation coefficient estimates the linear association between the predicted biomass and reference data expected at any point in the sampled population. However, many remotely sensed data are classified into categories.

Categorical data (*e.g.*, forest and nonforest cover) are assessed with the error-matrix, which is a contingency table cross-classified by reference and predicted categories. Count data from a simple random sample of point-plots support unbiased estimators of the proportion of the sampled population in each cell of the error-matrix (Congalton and Green 2009).

A statistical estimator is unbiased if the mean value of an estimated statistic among all possible samples is identical to the true population parameter. In practice, only one sample is realized. The resulting estimate surely deviates from the unknown population parameter. However, as the sample size becomes larger, the expected deviation becomes smaller, which reduces risk of incorrect inference. Regrettably, an adequate sample size is expensive with a simple random sample, especially for categorical data with a detailed taxonomic scheme or rare taxa.

A. Complex Sampling Designs

Complex sample designs can provide more reference data at less cost. For example, a field crew can measure ground reference data for a 1-ha cluster-plot with little incremental cost relative to measuring a single point-plot (*e.g.* a 1-m^2 "point"). Remotely sensed classifications of land cover for a contiguous cluster of pixels may be registered to the corresponding cluster-plot measured in the field. Assuming a simple random sample of equal-size cluster-plots, estimators with raw count data given by Congalton and Green (2009) remain unbiased for an error matrix. However, their variance estimators assume point-plots, and those estimators produce biased variance estimates when applied to raw count data from cluster-plots. A biased variance estimator constructs confidence intervals that do not faithfully cover the expected population parameters at the professed probability level. Any conclusions based on those confidence intervals can be unintentionally misleading.

Rare categories are often important (*e.g.*, old-growth forest habitat). Simple random sampling will rarely sample a rare category, and some rare categories can be entirely missing from a sample. Pre-stratification can prescribe a reasonable sample size for every rare category identified in the geospatial database. Furthermore, pre-stratification can use a higher sampling rate for rare and otherwise important categories.

A pre-stratified sampling design also produces raw count data for categorical variables, and the estimators in Congalton and Green (2009) appear applicable. However, those methods produce biased estimators of every proportion in the error-matrix in addition to misleading confidence intervals. As differences in sampling rates among strata become extreme, the bias becomes more serious (Czaplewski 2003). An unbiased estimator requires differentially weighting raw count data to adjust for unequal sampling rates (*e.g.*, Stehman and Foody 2009). It is seductively easy to ignore these weights and unwittingly produce a misleading accuracy assessment.

B. Key Lesson

Here is the key lesson. A statistical estimator can be design-unbiased when applied to a valid probability sample. However, that same estimator can be biased whenever applied to a sample selected with a different sampling design or one that uses different sampling units.

A biased estimator compromises strong inference regarding the accuracy of a geospatial database. The bias might be inconsequential, but the degree of bias is often unknown. Inference based on a biased estimator requires unsubstantiated assumptions.

There can be little difference in the cost of using a simple but biased estimator compared to a more complex but unbiased estimator. The next section briefly illustrates relatively simple estimators for a few complex designs.

II. Recursive Restriction Estimation

Knottnerus (2003 §12.2.2) introduced the General Restriction (GR) estimator into the sample survey literature. The GR estimator is closely related to the static Kalman filter (Maybeck 1979). In this setting, GR is a multivariate composite estimator that combines two separate sample survey components.

Knottnerus (2003 §12.5) also introduced the Recursive Restriction (RR) estimator, which extends the GR estimator into more complex applications. The RR estimator separates an even more complex design into a sequence of simpler estimators. These may include any standard sample survey estimator (*e.g.*, Särndal *et al.* 1992) and intermediate results with the GR estimator (Czaplewski 2010).

The example in the next section uses an accuracy assessment of airborne *Light Detection And Ranging* (LiDAR) laser data as predictors of above ground forest biomass. Field measurements from a *National Forest Inventory* (NFI) provide ground reference data (*e.g.*, McRoberts *et al*. 2005). The accuracy assessment statistic is the correlation coefficient between the predictions of biomass with LiDAR data and NFI field measurements.

In this example, a second independent survey adds a large sample of airborne LiDAR data, without the additional cost of NFI field data. The GR estimator improves the estimated correlation coefficient by combining results from the NFI survey with results from this second auxiliary LiDAR survey.

Finally, the example posits a third independent survey with a completely different set of sample plots. Spaceborne and airborne LiDAR sensor data predict biomass for each plot. The RR estimator combines results from the GR estimator with results from this third auxiliary survey.

A. Example of the General Restriction (GR) Estimator

Consider an accuracy assessment based on the correlation between remotely sensed estimates of forest biomass and corresponding ground reference data. Measurements for a relatively small, simple random sample of 1-ha NFI plots provide reference data. In addition, an airborne LiDAR sensor acquires data that predict biomass at each NFI field plot.

Each plot (i) has a 3×1 measurement vector: the first element is a function of the LiDAR data (X); the second element is a function of NFI field measurements (Y); and the third element is a function of their cross-product. A simple random sample, with a sample size of n_{NFI} plots, produces sufficient statistics for a design-consistent estimator (1) of the linear correlation coefficient (r_{NFI}) for the sampled population.

$$\hat{r}_{\text{NFI}} = \frac{\overline{\left(X-\bar{X}\right)\left(Y-\bar{Y}\right)}_{\text{NFI}}}{\sqrt{\overline{\left(X-\bar{X}\right)^2}_{\text{NFI}}\ \overline{\left(Y-\bar{Y}\right)^2}_{\text{NFI}}}} \quad \text{where}$$

$$\begin{bmatrix} \overline{\left(X-\bar{X}\right)^2}_{\text{NFI}} \\ \overline{\left(Y-\bar{Y}\right)^2}_{\text{NFI}} \\ \overline{\left(X-\bar{X}\right)\left(Y-\bar{Y}\right)}_{\text{NFI}} \end{bmatrix} = \left(\frac{1}{n_{\text{NFI}}-1}\right)\sum_{i=1}^{n}\begin{bmatrix} \left(X_i-\hat{\bar{X}}_{\text{NFI}}\right)^2 \\ \left(Y_i-\hat{\bar{Y}}_{\text{NFI}}\right)^2 \\ \left(X_i-\hat{\bar{X}}_{\text{NFI}}\right)\left(Y_i-\hat{\bar{Y}}_{\text{NFI}}\right) \end{bmatrix}$$

$$\begin{bmatrix} \hat{\bar{X}}_{\text{NFI}} \\ \hat{\bar{Y}}_{\text{NFI}} \end{bmatrix} = \left(\frac{1}{n_{\text{NFI}}}\right)\sum_{i=1}^{n}\begin{bmatrix} X_i \\ Y_i \end{bmatrix}$$

$$\mathbf{V}_{\text{NFI}} = \left\{ \frac{\sum_{i=1}^{n}\begin{bmatrix} \left(X_i-\hat{\bar{X}}_{\text{NFI}}\right)^2 \\ \left(Y_i-\hat{\bar{Y}}_{\text{NFI}}\right)^2 \\ \left(X_i-\hat{\bar{X}}_{\text{NFI}}\right)\left(Y_i-\hat{\bar{Y}}_{\text{NFI}}\right) \end{bmatrix}\begin{bmatrix} \left(X_i-\hat{\bar{X}}_{\text{NFI}}\right)^2 \\ \left(Y_i-\hat{\bar{Y}}_{\text{NFI}}\right)^2 \\ \left(X_i-\hat{\bar{X}}_{\text{NFI}}\right)\left(Y_i-\hat{\bar{Y}}_{\text{NFI}}\right) \end{bmatrix}'}{n_{\text{NFI}}\left(n_{\text{NFI}}-1\right)} \right\} \quad (1)$$

A linear Taylor-series approximation (Czaplewski 2010) provides a variance estimate for the correlation coefficient (r_{NFI}) using the 3×3 covariance matrix ($\mathbf{V}_{\text{NFI}}$) for the 3×1 vector estimate in (1). However, the sample size (n_{NFI}) of NFI plots is small, the estimated variance of the correlation coefficient is large, and the corresponding confidence interval is unacceptably broad in (1).

How can a more accurate estimate the correlation coefficient be made without a substantial increase in the cost of ground reference data? One answer is the addition of ancillary accuracy assessment data gathered with a less expensive protocol.

Assume auxiliary LiDAR data are acquired for a second survey that uses a simple random sample. The sample size can be large because LiDAR data are relatively inexpensive, and this survey does not include expensive field data. This second survey provides an independent unbiased estimator for the first element in the 3×1 vector estimate in (1).

$$\overline{\overline{(X-\bar{X})^2}}_{\text{LiDAR}} = \frac{\sum_{j=1}^{n}\left(X_i - \hat{\bar{X}}_{\text{LiDAR}}\right)^2}{n_{\text{LiDAR}}-1} \quad \text{where}$$

$$\hat{\bar{X}}_{\text{LiDAR}} = \frac{\sum_{j=1}^{n}[X_i]}{n_{\text{LiDAR}}} \qquad (2)$$

$$\left(\hat{v}_{(X-\bar{X})^2}\right)_{\text{LiDAR}} = \frac{\sum_{j=1}^{n}\left[\left(X_i - \hat{\bar{X}}_{\text{LiDAR}}\right)^2 - \overline{\overline{(X-\bar{X})^2}}_{\text{LiDAR}}\right]^2}{n_{\text{LiDAR}}\left(n_{\text{LiDAR}}-1\right)}$$

A multivariate version of Knottnerus' GR estimator combines these two independent surveys. The 3×1 vector estimate from the NFI survey (1) is weighted by a 3×3 matrix ($\mathbf{I}-\mathbf{k}_{\text{GR}}\mathbf{h}$), the scalar LiDAR estimate from the second survey (2) is weighted by a 3×1 vector $\mathbf{k}_{\text{GR}}$, and the two are summed into a single 3×1 GR estimate (3).

$$\hat{r}_{\text{GR}} = \frac{\overline{\overline{(X-\bar{X})(Y-\bar{Y})}}_{\text{GR}}}{\sqrt{\overline{\overline{(X-\bar{X})^2}}_{\text{GR}}\;\overline{\overline{(Y-\bar{Y})^2}}_{\text{GR}}}} \quad \text{where}$$

$$\begin{bmatrix} \overline{\overline{(X-\bar{X})^2}}_{\text{GR}} \\ \overline{\overline{(Y-\bar{Y})^2}}_{\text{GR}} \\ \overline{\overline{(X-\bar{X})(Y-\bar{Y})}}_{\text{GR}} \end{bmatrix} = \begin{Bmatrix} (\mathbf{I}-\mathbf{k}_{\text{GR}}\mathbf{h}) \begin{bmatrix} \overline{\overline{(X-\bar{X})^2}}_{\text{NFI}} \\ \overline{\overline{(Y-\bar{Y})^2}}_{\text{NFI}} \\ \overline{\overline{(X-\bar{X})(Y-\bar{Y})}}_{\text{NFI}} \end{bmatrix} \\ +\mathbf{k}_{\text{GR}}\overline{\overline{(X-\bar{X})^2}}_{\text{LiDAR}} \end{Bmatrix}$$

$$\hat{\mathbf{V}}_{\text{GR}} = \begin{Bmatrix} \left[(\mathbf{I}-\mathbf{k}_{\text{GR}}\mathbf{h})\hat{\mathbf{V}}_{\text{NRI}}(\mathbf{I}-\mathbf{k}_{\text{GR}}\mathbf{h})'\right] \\ +\left[\mathbf{k}_{\text{GR}}\left(\hat{v}_{(X-\bar{X})^2}\right)_{\text{LiDAR}}\mathbf{k}'_{\text{GR}}\right] \end{Bmatrix}$$

$$\mathbf{k}_{\text{GR}} = \left\{\frac{\hat{\mathbf{V}}_{\text{NRI}}\,\mathbf{h}'}{\mathbf{h}\hat{\mathbf{V}}_{\text{NFI}}\mathbf{h}' + \left(\hat{v}_{(X-\bar{X})^2}\right)_{\text{LiDAR}}}\right\} \qquad \mathbf{h} = [1 \quad 0 \quad 0] \qquad (3)$$

$\mathbf{V}_{\text{NFI}}$ is the 3×3 covariance matrix for the NFI vector estimate in (1), v_{LiDAR} is the scalar variance for the LiDAR biomass estimate from the second independent sample survey in (2), $\mathbf{h}$ is a 1×3 indicator vector, and $\mathbf{I}$ is the 3×3 identity matrix. The weights in the GR estimator ($\mathbf{k}_{\text{GR}}$) use the minimum variance optimality criterion (Maybeck 1979 p. 232). Improved estimates of the first vector element will improve precision of the remaining two elements depending on the covariances among all three elements ($\mathbf{V}_{\text{NFI}}$) from (1). Therefore, the estimated correlation coefficient with auxiliary LiDAR data (r_{GR}) in (3) is expected to have a smaller confidence interval than r_{NFI} in (1), which uses data from NFI field plots alone.

B. Example of the Recursive Restriction (RR) Estimator

The Recursive Restriction (RR) estimator, which is a simple extension of the GR estimator, accommodates designs that are more complex. For example, assume broad swaths of spaceborne LiDAR data from the Shuttle Radar Topographic Mission (SRTM) provide a systematic sample of transects for the sampled population. Airborne LiDAR data are acquired within each transect. This third independent sample[1] provides additional auxiliary data. The sequential RR estimator uses the SRTM data to improve precision of the first element in the GR vector estimator (3).

$$\hat{r}_{\text{RR}} = \frac{\overline{\overline{(X-\bar{X})(Y-\bar{Y})}}_{\text{RR}}}{\sqrt{\overline{\overline{(X-\bar{X})^2}}_{\text{RR}}\;\overline{\overline{(Y-\bar{Y})^2}}_{\text{RR}}}} \quad \text{where}$$

$$\begin{bmatrix} \overline{\overline{(X-\bar{X})^2}}_{\text{RR}} \\ \overline{\overline{(Y-\bar{Y})^2}}_{\text{RR}} \\ \overline{\overline{(X-\bar{X})(Y-\bar{Y})}}_{\text{RR}} \end{bmatrix} = \begin{Bmatrix} (\mathbf{I}-\mathbf{k}_{\text{RR}}\mathbf{h}) \begin{bmatrix} \overline{\overline{(X-\bar{X})^2}}_{\text{GR}} \\ \overline{\overline{(Y-\bar{Y})^2}}_{\text{GR}} \\ \overline{\overline{(X-\bar{X})(Y-\bar{Y})}}_{\text{GR}} \end{bmatrix} \\ +\mathbf{k}_{\text{RR}}\left(\overline{\overline{(X-\bar{X})^2}}_{\text{SRTM}}\right) \end{Bmatrix}$$

$$\hat{\mathbf{V}}_{\text{RR}} = \begin{Bmatrix} \left[(\mathbf{I}-\mathbf{k}_{\text{RR}}\mathbf{h})\hat{\mathbf{V}}_{\text{GR}}(\mathbf{I}-\mathbf{k}_{\text{RR}}\mathbf{h})'\right] \\ +\left[\mathbf{k}_{\text{RR}}\left(\hat{v}_{(X-\bar{X})^2}\right)_{\text{SRTM}}\mathbf{k}'_{\text{RR}}\right] \end{Bmatrix}$$

$$\mathbf{k}_{\text{RR}} = \left\{\frac{\hat{\mathbf{V}}_{\text{GR}}\,\mathbf{h}'}{\mathbf{h}\hat{\mathbf{V}}_{\text{GR}}\mathbf{h}' + \left(\hat{v}_{(X-\bar{X})^2}\right)_{\text{SRTM}}}\right\} \qquad (4)$$

$$\overline{\overline{(X-\bar{X})^2}}_{\text{SRTM}} = \frac{\sum_{j=1}^{n}\left(X_i - \hat{\bar{X}}_{\text{SRTM}}\right)^2}{n_{\text{SRTM}}-1} \qquad \hat{\bar{X}}_{\text{SRTM}} = \frac{\sum_{j=1}^{n}[X_i]}{n_{\text{SRTM}}}$$

$$\left(\hat{v}_{(X-\bar{X})^2}\right)_{\text{SRTM}} = \frac{\sum_{j=1}^{n}\left[\left(X_i - \hat{\bar{X}}_{\text{SRTM}}\right)^2 - \overline{\overline{(X-\bar{X})^2}}_{\text{SRTM}}\right]^2}{n_{\text{SRTM}}\left(n_{\text{SRTM}}-1\right)}$$

The estimated correlation coefficient r_{RR} in (4) is expected to have a smaller confidence interval than r_{GE} in (3) because the SRTM data improves all elements in the 3×1 vector estimates used to compute r_{RR}.

[1] For the sake of simplicity, the sampling design with spaceborne SRTM LiDAR data is treated as a simple random sample in (4). See Czaplewski (in press) for methods that are more efficient and suitable to two-stage sampling with SRTM transects. These methods improve the estimate of biomass with the airborne LiDAR with auxiliary data from the spaceborne SRTM LiDAR.

III. DISCUSSION

More complicated sample surveys can improve cost-effectiveness (Särndal *et al.* 1992). Pre-stratification for rare or important categories, which served as an example in the Introduction, is one example.

Another example is an accuracy assessment of biomass predictions with spaceborne Landsat sensor data. Those predictions can be assessed with a large sample of 100-ha cluster-plots. Each cluster-plot may be measured with airborne LiDAR and/or photogrammetric measurements with aerial photographs, and a small sub-sample can be further measured in the field.

The sampled population may be stratified by accessibility to more efficiently allocate field sampling. Two or more independent surveys with compatible measurement protocols can be optimized for special objectives or different circumstances.

Different configurations for sample plots can be optimized for each sensor. These may be nested together into multi-stage designs that collate predictions from different sensors. For example, Frescino *et al*. (2009) use a 1-ha NFI field plot nested within a 50-ha plot that is designed for very high-resolution large-scale aerial photography. However, the GR and RR estimators apply to more complex designs. Consider a full-coverage global geospatial database based on the spaceborne MODIS sensor. The accuracy assessment might use a nested sampling unit, where a 10^6-ha Landsat scene is the Primary Sampling Unit. Several 10^4-ha Secondary Sampling Units within each sampled Landsat scene are measured with the spaceborne IKONOS sensor. Each of those is sub-sampled with several 10^2-ha Tertiary Sampling Units measured with large-scale aerial photography, one of which covers a 1-ha NFI field plot as the Quaternary Sampling Unit.

These multi-stage components are not mutually independent because sampling units are collocated. Czaplewski (2010) extends the multivariate RR estimator to multi-stage designs. He uses results from Maybeck (1979 p. 247) that accommodate nonzero covariances among otherwise separable design components. Similar multivariate RR estimation methods provide relatively simple and efficient design-consistent estimators that blend multi-phase and multi-stage components[2].

Error matrices estimated with a complex sampling design require applicable multivariate methods to estimate variances, standard deviations and confidence intervals. Czaplewski (1994) derives generalized variance estimators for common accuracy assessment statistics. These include conditional probabilities, such as users' and producers' accuracies and the kappa statistic.

[2] Czaplewski (2010) discusses the relationship between multivariate GR and RR estimators and univariate calibration and regression estimators (Särndal *et al.* 1992).

IV. CONCLUSION

Complex sampling designs offer promising opportunities to improve accuracy assessments, especially within large remote sensing projects. However, complex methods incur some risk. An expert statistician can minimize that risk. Knottnerus' GR and RR estimators provide that expert with relatively simple, flexible, efficient and unbiased estimators for any complex sampling design.

ACKNOWLEDGEMENTS

The following reviewers substantially helped improve the presentation of these concepts: Stephen Stehman, Giles Foody, Timothy Gregoire, and Lyman McDonald. Any errors or omissions remain the sole responsibility of the author.

REFERENCES

Congalton, R. G. and Green, K. (2009). *Assessing the accuracy of remotely sensed data: Principles and practices*, 2nd ed., Boca Ralton, FL: CRC Press.

Czaplewski, R. L. (1994). *Variance approximations for assessments of classification accuracy*. Research Paper RM-316, Fort Collins, CO: United States Forest Service, Rocky Mountain Research Station.

Czaplewski, R. L. (2003). *Accuracy assessment of maps of forest condition: statistical design and methodological considerations*. In M. A. Wulder and S. E. Franklin (Eds) *Remote Sensing of Forest Environments: Concepts and Case Studies* (pp. 115-140). Boston, MA: Kluwer Academic Publishers.

Czaplewski, R. L. (2010). *Complex sample survey estimation in static state-space*, General Technical Report RM-239, Fort Collins, CO: United States Forest Service, Rocky Mountain Research Station.

Frescino, T., Moisen G.G., Megown, K., Nelson, V., Freeman, E., Patterson, P., Finco, M., Brewer, K., and Menlove, J. (2009) Nevada Photo-Based Inventory Pilot (NPIP) photo sampling procedures. In W.McWilliams, G.G. Moisen and R.L. Czaplewski (Eds) *Proceedings of the 2008 Forest Inventory and Analysis (FIA) Symposium*. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 1-30 p.

Knottnerus, P. (2003) *Sample survey theory: Some Pythagorean perspectives*, New York: Springer-Verlag.

Maybeck, P. S. (1979) *Stochastic models, estimation, and control*, vol. 141-1 of *Mathematics in science and engineering*, New York: Academic Press.

McRoberts, R.E., Bechtold, W.A., Patterson, P.L., Scott, C.T., Reams, G.A. (2005). The enhanced Forest Inventory and Analysis program of the USDA Forest Service: Historical perspective and announcement of statistical documentation, *Journal of Forestry* 103(6), 304-308.

Särndal, C. E., Swensson, B., and Wretman, J. H. (1992). *Model assisted survey sampling*, New York: Springer-Verlag.

Stehman, S. V. and Foody, G. M. (2009). *Accuracy assessment*. In T. A. Warner, M. D. Nellis, and G. M. Foody (Eds) *The SAGE Handbook of Remote Sensing* (pp. 297-308). London: SAGE Publications.