Misclassification Bias in Areal Estimates

Raymond L. Czaplewski
USDA Forest Service, Rocky Mountain Forest and Range Experiment Station, 240 W. Prospect Street, Fort Collins, CO 80526–2098

ABSTRACT: In addition to thematic maps, remote sensing provides estimates of area in different thematic categories. Areal estimates are frequently used for resource inventories, management planning, and assessment analyses. Misclassification causes bias in these statistical areal estimates. For example, if a small percentage of a common cover type is misclassified as a rare cover type, then the area occupied by the rare type can be severely overestimated. Many categories are rare in detailed classification systems. I present an informal method to anticipate the approximate magnitude of this bias in statistical areal estimates, before a remote sensing study is conducted. If the anticipated magnitude is unacceptable, then statistical calibration methods should be used to produce unbiased areal estimates. I then discuss existing statistical methods that calibrate for misclassification bias with a sample of reference plots.

INTRODUCTION

REMOTELY SENSED AREAL ESTIMATES are typically treated as unbiased estimates of the true area for each cover type in a study area. However, Card (1982), Chrisman (1982), and Hay (1988) note that misclassification can bias areal estimates from remote sensing. My first objective is to demonstrate the cause of misclassification bias, and then to present an informal method to anticipate its magnitude before a remote sensing study is conducted. With a quantitative expectation of the approximate magnitude of this bias, the user of remotely sensed areal estimates can judge the practical importance of misclassification bias, given the unique requirements of each remote sensing study. If the anticipated magnitude of misclassification bias is unacceptable, then areal estimates should be calibrated with remotely sensed and reference classifications for a representative sample of reference plots from the study area. My second objective is to increase awareness of existing methods that can statistically calibrate for misclassification bias, and then to provide general guidance in the choice and application of an appropriate calibration method.

SOURCE OF MISCLASSIFICATION BIAS

Let the remotely sensed percentage of cover type A be denoted as the scalar Y, and the true percentage of cover type A be denoted as the scalar X. The true percentage of cover types other than A (labeled cover type B in the following) will equal the scalar $(100\% - X)$. Let scalar H_A represent the conditional probability that any pixel is interpreted as cover type A, given that the pixel is truly cover type A, where $0 \leq H_A \leq 1$; and let scalar $(1 - H_B)$ represent the conditional probability that any pixel is interpreted as cover type A, given it is truly type B, where $0 \leq H_B \leq 1$. H_A and H_B represent producer's accuracies, and are measures of omission error (Story and Congalton, 1986). The remotely sensed percentage (Y) of cover type A will be the following deterministic function of the true percentage (X) of cover type A and the true conditional probabilities of omission errors (H_A and H_B):

$$Y = [H_A X] + [(1 - H_B)(100\% - X)]. \quad (1)$$

Equation 1 shows that misclassification biases areal estimates from remote sensing; the remotely sensed percentage (Y) will not equal the true percentage (X) unless there are no omission errors, i.e., $H_A = H_B = 1$, or effects of omission errors exactly compensate, i.e., $(1 - H_B)(100\% - X) = (1 - H_A) X$. Either condition is rare in remote sensing.

Proportions or acreages of each cover type can be readily used in place of percentages in Equation 1. Instead of $(100\% - X)$, $(1 - X)$ would be used if X and Y are proportions, and $(T - X)$ would be used if X and Y are acreages, where T is the total acreage of the study area.

Assume classification accuracies are high for all cover types (e.g , $H_A = H_B = 0.95$). If cover type A truly occupies 90 percent of the study area (i.e., $X = 90$), then the remotely sensed percentage (Y) will equal 86 (see Equation 1). Similarly, Y equals 68 percent if X equals 70 percent, and Y equals 50 percent if X equals 50 percent. The bias in areal estimates for rare categories can be relatively high, even with such high classification accuracies. If cover type A truly occupies 10 percent of the study area, then the remotely sensed estimate will be 14 percent (see Equation 1). In this example, the remotely sensed percentage will be 40 percent larger than the true value. If a small percentage of a common cover type is misclassified as a rare cover type, then the area occupied by the rare type will be overestimated, unless there is a high rate of omission error in classifying the rare type. As the detail of a classification system increases, many categories will be rare. Figure 1 portrays the magnitude of misclassification bias for a wide range of classification accuracies.

MAGNITUDE OF MISCLASSIFICATION BIAS

Figure 1 or Equation 1 can be informally used to anticipate the approximate magnitude of misclassification bias for any cover type. However, preliminary expectations of classification accuracies and prevalence of various cover types must be used, rather than their true, but unknown, values. For example, assume you expect that classification accuracies for your study area will be similar to those given by Story and Congalton (1986), who used reference and remotely sensed classifications of 30 forested plots, 30 water plots, and 40 urban plots to construct an error matrix. From their error matrix, your preliminary estimate of producer's accuracy for the forest cover type in your study area is $H_A = 28/30 = 0.93$, and your preliminary estimate for non-forest accuracy is $H_B = (15+1+5+20)/(30+40) = 0.59$ (i.e., the water and urban types are pooled together). Assume your preliminary estimate of forest cover in your study area is 33 percent. Using Figure 1 with $H_A = 0.90$, $H_B = 0.60$, and $X = 33$, you anticipate misclassification bias will be approximately 25 percent; you can expect the remotely sensed areal estimate for forest will be $(33 + 25) = 58$ percent if forest cover is truly 33 percent in your study area. From the same error matrix, producer's accuracy for water cover is $H_A = 15/30 = 0.50$, and that for non-water (i.e., forest and urban) is $H_B = (28+1+15+20)/(30+40) = 0.91$. You can anticipate from Equation 1 that the remotely sensed areal estimate for water will be approximately $Y = 23$ percent if water truly covers $X = 33$ percent of your study area, i.e., misclassification bias of -10 percent. Finally, you can use Equation 1 and the same error matrix to anticipate that the remotely

PHOTOGRAMMETRIC ENGINEERING & REMOTE SENSING,
Vol. 58, No. 2, February 1992, pp. 189-192.

0099-1112/92/5802–189$03.00/0

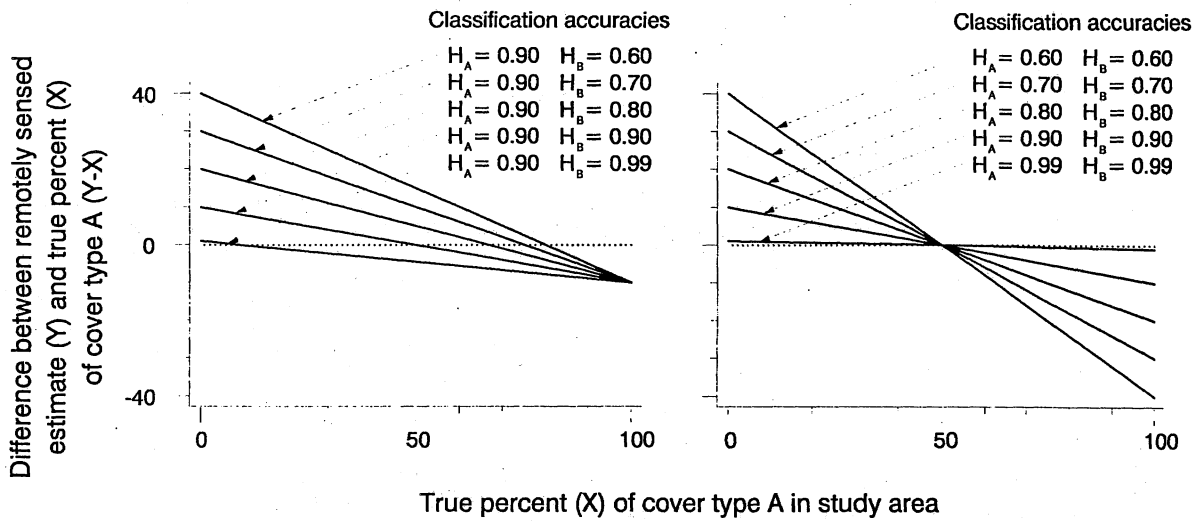

FIG. 1. Examples demonstrating magnitude of misclassification bias for various probabilities of omission errors. Misclassification bias is the difference between the remotely sensed estimate and the true percentage of a cover type. Remotely sensed estimate (Y) is a function (Equation 1) of prevalence of a cover type (X) and producer's classification accuracies (H_A and H_B), which are conditional probabilities of omission errors. H_A and H_B are sometimes equal in this example, but they are not necessarily equal in practice. This figure, or Equation 1, can be informally used to anticipate the magnitude of misclassification bias, given approximate expectations of classification accuracy and prevalence of cover types. If the anticipated magnitude is unacceptable, then more formal calibration methods should be considered, which are discussed in this paper.

sensed areal estimate for urban will be approximately 19 percent if urban areas truly cover 33 percent of your study area, i.e., misclassification bias of −14 percent.

CALIBRATION FOR MISCLASSIFICATION BIAS

If the anticipated magnitude of misclassification bias is unacceptable to the user, then more formal calibration techniques should be included in the remote sensing study. Calibration cannot identify misclassified pixels. Rather, calibration is a probabilistic technique; it uses proportions of imperfectly classified pixels in a reference sample to estimate conditional probabilities of various types of misclassification, and these estimated probabilities are then used to predict the true percentage of each cover type given the remotely sensed percentages. Proportions or acreages can be used in place of percentages. Statistical calibration requires accurate estimates of misclassification probabilities using reference plots from the study area, rather than preliminary expectations used above to anticipate the approximate magnitude of misclassification bias.

Two different calibration methods can be used if a reliable error matrix is acquired for a particular study. In remote sensing, Bauer *et al.* (1978), Maxim *et al.* (1981), Prisley and Smith (1987), and Hay (1988) demonstrate use of a classical multivariate calibration method, which was introduced into the statistical literature by Grassia and Sundberg (1982). Equation 1 is a univariate example of this method. To produce an unbiased areal estimate, Equation 1 is solved for the true percentage (X) given the remotely sensed estimate (Y), and accurate estimates of the probabilities of omission errors (i.e., H_A and H_B). The inverse calibration estimator of Tenenbein (1972) is an alternative to this classical estimator, and Card (1982) and Chrisman (1982) demonstrate use of this estimator in remote sensing. The inverse estimator uses probabilities of commission errors (i.e., user's accuracies), while the classical estimator uses probabilities of omission errors. Czaplewski and Catts (1990) give examples of these two calibration methods in remote sensing.

Based on unpublished Monte Carlo simulations, I found the inverse calibration estimator of Tenenbein is less biased, more precise, and less prone to numerical problems and infeasible solutions, especially for small sample sizes of reference plots. For example, the multivariate classical estimator requires a matrix inversion, and can produce negative areal estimates; the inverse estimator requires less complex algebra, and will always produce positive estimates.

These calibration techniques use misclassification probabilities from an error matrix that are estimated with a finite sample of reference plots from the study area. These estimated probabilities contain sampling errors, which are propagated into estimation errors for the true percentage of each category in the study area. As the sample size of reference plots increases, the sampling error deceases for estimates of misclassification probabilities, and accuracy of the calibrated areal estimate increases. Grassia and Sundberg (1982) and Tenenbein (1972) give approximate covariance matrices for these estimation errors; they assume a large sample of reference plots is available, and the reference plots are independent and homogeneous (i.e., each independent reference plot is classified into a single category with remote sensing, and a single category with the reference data). These covariance matrices are needed to construct confidence intervals, which describe the level of uncertainty in the calibrated areal estimate that is produced by uncertain estimates of misclassification probabilities.

An unstratified sample of homogeneous reference plots will include a small number of rare cover types. Stratification can provide more intensive sampling of rare types, which can improve accuracy of calibrated areal estimates for rare types. However, an inappropriate calibration technique can bias calibrated areal estimates from a stratified sample of reference plots. In general, the inverse estimator of Tenenbein should be used if stratification is based on the remotely sensed classifications. If the stratified sample is selected based on the reference classifications, then the classical estimator of Grassia and Sundberg should be used. This latter situation might exist if existing field plots are used for reference data, but the cost of accurate registration of existing field plots to the remotely sensed imagery limits the number of plots that can be registered. Bias from an inappropriate calibration technique can be eliminated with independent ancillary estimates of the true or remotely sensed

percentages in the study area, but more elaborate calibration methods are required.

These calibration methods require a large sample of representative reference plots to estimate misclassification probabilities. Representative reference plots are best selected with randomization methods. Training or labeling plots often have lower rates of classification error than are typical for the entire study area, and such plots will produce biased areal estimates if used for calibration. Brown (1982) discusses controlled calibration, which can use purposefully selected reference plots; however, this requires Bayesian estimation, which is vulnerable to subtle problems and undetected biases.

If the reference plots are not well registered to the remotely sensed imagery, then classification error will be confounded with registration error, and the misclassification probabilities will be poorly estimated. Large heterogeneous reference plots might be more successfully registered to remote sensing imagery than small homogeneous plots. If the heterogeneous plots are a simple random or systematic sample, and reference classifications are available for each pixel in these plots, then the calibration methods of Tenenbein (1972) and Grassia and Sundberg (1982) can be applied without modification. However, classification errors for adjacent pixels in the same reference plot are not independent, and different methods would be required to calculate the estimation error covariance matrix.

Different calibration methods are required if an error matrix cannot be constructed from the reference data. For example, reference data for agricultural surveys can be limited to areal estimates of different crop covers within large, heterogeneous plots; maps showing the location of each crop cover within the reference plots might not be available. Therefore, the reference classifications for each pixel within the reference plots are not available, and an error matrix cannot be constructed. Here, calibration can only use the remotely sensed and reference percentages of each cover type within each heterogeneous reference plot. Calibration estimators for this situation have been developed and evaluated by Chhikara *et al.* (1986), Fuller (1986), Heydorn and Takacs (1986), McKeon and Chhikara (1986), Hung and Fuller (1987), Battese *et al.* (1988), and Chhikara and Deng (1988). Similar situations arise when registration of pixels to reference plots is problematic, and reference classifications for individual pixels cannot be reliably obtained. Iverson *et al.* (1989) consider this situation for AVHRR data, where remotely sensed estimates for Landsat scenes serve as the reference data. In addition, Pech *et al.* (1986) describe a method to calibrate areal estimates from mixed pixels that cannot be classified into unique categories. All these methods are linear regression techniques rather than the probabilistic techniques of Tenenbein (1972) and Grassia and Sundberg (1982). Calibration based on regression methods can produce negative areal estimates. Lewis and Odell (1971), Liew (1976), and Shim (1983) propose quadratic programming techniques to avoid negative estimates, and van Roessel (USDA Forest Service, 1980) has applied this solution in remote sensing. Detection limits can affect misclassification bias in more complex ways. For example, an AVHRR pixel might require 30 percent deforestation before any deforestation can be detected. This can cause a nonlinear relationship between the remotely sensed and reference areal estimates for the reference plots, which might require nonlinear calibration estimators. Scheffé (1973) and Brown (1982) discuss the statistical aspects of nonlinear calibration, but this technique has not been applied in remote sensing.

All of these calibration methods correct areal estimates for misclassification error, despite the cause. Interpretation error is the most familiar cause. However, changes in land cover might occur between the dates that remotely sensed images and reference data were acquired, or there might be differences in definitions between the remote sensing and reference classification systems. Calibration treats the reference data as the standard, and calibrated areal estimates represent the acquisition dates definitions and protocol used for the reference data. For example, if the remotely sensed images were acquired in 1987 and the reference data in 1991, then the calibrated areal estimates are an unbiased estimate of the status in 1991. If users require areal estimates consistent with their existing definitions and protocol for field surveys, but other methods are used for the reference data in calibration (e.g., photointerpretation of large-scale imagery, or "windshield surveys"), then the calibrated areal estimates can be unacceptable to the user.

All of these calibration techniques are closely related to various multi-stage or multi-phase sampling designs, which can be more efficient than calibration if the sample size of reference plots is large. The remotely sensed data are analogous to the first level of a multi-level design, and the reference data are analogous to the second level. However, calibration methods have been developed that use areal estimates from all pixels in an image, and for multivariate and nonlinear situations; calibration might be more readily applied to these more complicated estimation problems than multi-level sampling designs.

CONCLUSIONS

Some users reject areal estimates from remote sensing because the magnitude of misclassification bias might be large. Some remote sensing specialists recommend that users ignore misclassification bias if classification accuracy is high. The most reasonable alternative might lay between these extremes. During the planning stage, remote sensing specialists should anticipate the approximate magnitude of misclassification bias. If the anticipated magnitude is unacceptable to the user of remotely sensed areal estimates, then the study plan should require statistical methods that will calibrate the final areal estimates. Reliable calibration requires an adequate, representative, and timely sample of accurately registered reference data from the study area.

REFERENCES

Battese, G. E., R. M. Harter, and W. A. Fuller, 1988. An error-component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*. 83:28–36.

Bauer, M. E., M. M. Hixson, B. J. Davis, and J. B. Etheridge, 1978. Area estimation of crops by digital analysis of Landsat data. *Photogrammetric Engineering & Remote Sensing*. 44:1033–1043.

Brown, P. J., 1982. Multivariate calibration. *Journal of the Royal Statistical Society, B*. 3:287–321.

Card, D. H., 1982. Using known map categorical marginal frequencies to improve estimates of thematic map accuracy. *Photogrammetric Engineering & Remote Sensing*. 48:431–439.

Chhikara, R. S., J. C. Lundgren, and A. A. Houston, 1986. Crop acreage estimation using a Landsat-based estimator as an auxiliary variable. *International Transactions in Geoscience and Remote Sensing*. 24:157–168.

Chhikara, R. S., and L. Y. Deng, 1988. Conditional inference in finite population sampling under a calibration model. *Communications in Statistics - Simulation and Computation*. 17:663–681.

Chrisman, N. R., 1982. Beyond accuracy assessment: correction of misclassification. *Proceedings of the 5th International Symposium on Computer-assisted Cartography*, Crystal City, Virginia. pp.123–132.

Czaplewski, R. L., and G. P. Catts, 1990. Calibrating area estimates for classification error using confusion matrices. *Proceedings of the 56th Annual Meeting of the American Society for Photogrammetry and Remote Sensing*, Denver, Colorado, Volume 4. pp. 431–440.

Fuller, W. S., 1986. Small area estimation as a measurement error problem. *Proceedings of the Conference on Survey Research Methods in Agriculture*, Leesburg, Virginia, 15–18 June.

Grassia, A., and R. Sundberg, 1982. Statistical precision in the calibra-

tion and use of sorting machines and other classifiers. *Technometrics*. 24:117–121.

Hay, A. M., 1988. The derivation of global estimates from a confusion matrix. *International Journal of Remote Sensing*. 9:1395–1398.

Heydorn, R. P., and H. C. Takacs, 1986. On the design of classifiers for crop inventories, *IEEE Transactions in Geoscience and Remote Sensing*. 24:150–155.

Hung, H. M., and W. A. Fuller, 1987. Regression estimation of crop acreages with transformed Landsat data as auxiliary variables. *Journal of Business and Economic Statistics*. 5: 475–482.

Lewis, T. O., and P. L. Odell, 1971. *Estimation in Linear Models*. Prentice-Hall, New Jersey, 193 p.

Liew, C. K., 1976. Inequality constrained least-squares estimation. *Journal of the American Statistical Association*. 71:746–751.

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 Journal of Remote Sensing*. 10:1805–1812.

Maxim, L. D., L. Harrington, and M. Kennedy, 1981. Alternative scale-up estimates for aerial surveys where both detection and classification error exist. *Photogrammetric Engineering & Remote Sensing*. 47:1227–1239.

McKeon, J. J., and R. S. Chhikara, 1986. Crop acreage estimation using satellite data as auxiliary information: multivariate case. *Proceedings of the Survey Research Methods Section, American Statistical Association*, University of Houston, Clear Lake, Texas.

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*. 7:1729–1750.

Prisley, S. P., and J. L. Smith, 1987. Using classification error matrices to improve the accuracy of weighted land-cover models. *Photogrammetric Engineering & Remote Sensing* 53:1259–1263.

Scheffé H., 1973. A statistical theory of calibration. *American Statistician*. 1:1–37.

Shim, J. K., 1983. A survey of quadratic programming applications to business and economics, *International Journal of Systems Science*. 14:105–115.

Story, M., and R. G. Congalton, 1986. Accuracy assessment; a user's perspective. *Photogrammetric Engineering & Remote Sensing*. 52:397–399.

Tenenbein, A., 1972. A double sampling scheme for estimating from misclassified multinomial data with applications to sampling inspection. *Technometrics* 14:187–202.

USDA Forest Service, 1980. *Evaluation of Multiresource Analysis and Information System (MAIS) Processing Components, Kershaw County South Carolina Feasibility Test*, Berkeley, California. 96 p.

(Received 27 June 1990; revised and accepted 19 March 1991)

Forthcoming Articles

Paul V. Bolstad and *T. M. Lillesand,* Rule-Based Classification Models: Flexible Integration of Satellite Imagery and Thematic Spatial Data.

John Crews, Overplotting Digital Geographic Data onto Existing Maps.

Claude R. Duguay and *Ellsworth F. LeDrew,* Estimating Surface Reflectance and Albedo from Landsat-5 Thematic Mapper over Rugged Terrain.

Peter F. Fisher, First Experiments in Viewshed Uncertainty: Simulating Fuzzy Viewsheds.

Steven E. Franklin and *Bradley A. Wilson,* A Three-Stage Classifier for Remote Sensing of Mountain Environments.

Clive S. Fraser, Photogrammetric Measurement to One Part in a Million.

Clive S. Fraser and *James A. Mallison,* Dimensional Characterization of a Large Aircraft Structure by Photogrammetry.

Clive S. Fraser and *Mark R. Shortis,* Variation of Distortion within the Photographic Field.

Peng Gong and *Philip J. Howarth,* Frequency-Based Contextual Classification and Gray-Level Vector Reduction for Land-Use Identification.

Christian Heipke, A Global Approach for Least-Squares Image Matching and Surface Reconstruction in Object Space.

David P. Lanter and *Howard Veregin,* A Research Paradigm for Propagating Error in Layer-Based GIS.

Richard G. Lathrop, Jr., Landsat Thematic Mapper Monitoring of Turbid Inland Water Quality.

Kurt Novak, Rectification of Digital Imagery.

J. Olaleye and *W. Faig,* Reducing the Registration Time for Photographs with Non-Intersecting Crossarm Fiducials on the Analytical Plotter.

Albert J. Peters, Bradley C. Reed, and *Donald C. Rundquist,* A Technique for Processing NOAA)AVHRR Data into a Geographically Referenced Image Map.

Kevin P. Price, David A. Pyke, and *Lloyd Mendes,* Shrub Dieback in a Semiarid Ecosystem: The Integration of Remote Sensing and Geographic Information Systems for Detecting Vegetation Change.

Benoit Rivard and *Raymond E. Arvidson,* Utility of Imaging Spectrometry for Lithologic Mapping in Greenland.

Kathryn Connors Sasowsky, Gary W. Petersen, and *Barry M. Evans,* Accuracy of SPOT Digital Elevation Model and Derivatives: Utility for Alaska's North Slope.

Omar H. Shemdin and *H. Minh Tran,* Measuring Short Surface Waves with Stereophotography.

Vittala K. Shettigara, A Generalized Component Substitution Technique for Spatial Enhancement of Multispectral Images Using a Higher Resolution Data Set.

Michael B. Smith and *Mitja Brilly,* Automated Grid Element Ordering for GIS-Based Overland Flow Modeling.

David M. Stoms, Frank W. Davis, and *Christopher B. Cogan,* Sensitivity of Wildlife Habitat Models to Uncertainties in GIS Data.

Khagendra Thapa and *John Bossler,* Accuracy of Spatial Data Used in Geographic Information Systems.

Thierry Toutin, Yves Carbonneau, and *Louiselle St-Laurent,* An Integrated Method to Rectify Airborne Radar Imagery Using DEM.

Paul M. Treitz, Philip J. Howarth, and *Peng Gong,* Application of Satellite and GIS Technologies for Land-Cover and Land-Use Mapping at the Rural-Urban Fringe: A Case Study.

William S. Warner and *Øystein Andersen,* Consequences of Enlarging Small-Format Imagery with a Color Copier.

Zhuoqiao Zeng and *Xibo Wang,* A General Solution of a Closed Form Space Resection.