Reprinted from

Seventh International Symposium

Machine Processing of

Remotely Sensed Data

with special emphasis on

Range, Forest and Wetlands Assessment

June 23 - 26, 1981

Proceedings

Purdue University
The Laboratory for Applications of Remote Sensing
West Lafayette, Indiana 47907 USA

TECHNIQUES FOR EVALUATION OF AREA ESTIMATES

MARILYN M. HIXSON

Purdue University/LARS
West Lafayette, Indiana

I. INTRODUCTION

Since the launch of the first Landsat satellite in 1972, satellite remote sensing has been increasingly recognized as a tool for mapping and area estimation of earth resources. The Landsat MSS records a region on the ground about one acre (0.5 ha) in size. This provides a good spatial resolution for mapping purposes, and Landsat data have been used for mapping such characteristics as general land use and soil type. Estimation of the areal extent of a feature has been a key use of Landsat data. The primary uses for area estimation have been in agriculture with crop and forest area estimation.

Although many researchers and users have analyzed Landsat data, the matter of determining and expressing in a meaningful and useful way the quality of a classification is a difficult problem. In evaluation of classification results, the experimenter may be concerned with two types of accuracy: classification accuracy and proportion estimation accuracy. By classification accuracy, we refer to the pixel-by-pixel count of the percentage of times the decision rule has produced the correct response. By proportion estimation accuracy, we refer to how close an estimate (e.g., of crop proportion) is to the "truth" or to some reference standard.

In the application of remote sensing technology to the problem of area estimation, classification accuracy may not be of prime importance. Compensating classification errors among categories or methods of estimation may enable the researcher to obtain accurate area estimates without attaining a classification accuracy as high as might be needed for mapping purposes.

Proportion estimates of classes of interest can be computed by direct estimation or unbiased estimation methods. The accuracy of these proportions can be assessed with respect to some reference standard or can be compared with results from other data analyses. This paper addresses methods of proportion estimation and qualitative and quantitative methods for evaluation of area or proportion estimates.

II. COMPONENTS OF QUALITY

In evaluation of a classification, two components of its quality must be evaluated: unbiasedness and precision.

By unbiasedness, we mean a low error rate. If X is an estimate found from a sample, the expected value of X is

$$E(X) = \sum_{R} RP(X = R)$$

where the sum extends over all the possible values R of X. X will be unbiased if E(X) is equal to the quantity being estimated.

The concept of precision refers to an estimate with a low variance. The variance of the estimate X is defined by

$$V(X) = E\left[X - E(X)\right]^2 .$$

The variance measures the amount of variation or scatter which would be observed among values of X, if the estimation procedure were conducted repetitively.

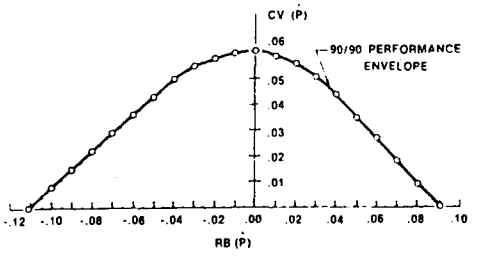

Figure 1. Relative bias and coefficients of variation which satisfy the 90/90 accuracy criterion.

The importance of both of the components of bias and precision can be illustrated by the Large Area Crop Inventory Experiment (LACIE) performance goal (1). The objective of LACIE was to satisfy the "90/90 criterion" for wheat production estimation; i.e., to be within 10% of the true proportion 90% of the time. Specifically, this criterion was to satisfy

$$\text{Prob}\left\{|\hat{P} - P| \leq 0.1\,P\right\} \geq 0.9$$

where P is the LACIE estimate of wheat production and P is the true wheat production. This criterion can be satisfied by a range of bias and precision values (Figure 1). It can be seen that the two are related in that larger biases can be permitted when estimates are very precise, for example.

III. EVALUATION OF AREA ESTIMATES

With the objective of evaluation of area estimates, the researcher may want to consider several types of measures. As in mapping, the classification accuracy should be examined; however, a pixel-by-pixel evalution is not sufficient to assess the quality of area estimates. The area or proportion estimates themselves must be evaluated, either by comparison with a reference standard or with results from another analysis.

Four specific areas will be addressed in this paper:

1. Estimation of the classification accuracy,
2. Estimation of proportions from classification results,
3. Comparison of area or proportion estimates with a reference standard, and
4. Comparison of area or proportion estimates with the results from another analysis.

A. CLASSIFICATION ACCURACY ESTIMATION

Selection of a Test Sample. Overall or cover-type specific classification accuracies are most generally estimated based on a set of test samples. A test sample can form a base for statistical evaluation if it is of sufficient size, represents all the variation present in the area, and has been selected using probability (random) sampling.

Selection of Sample Size. The estimation of sample size requires:

1. A required precision (d) that expresses how close to the true mean that the sample mean should be.
2. A measure of the variability in the population (σ^2).
3. A specification of the acceptable risk (α) that the actual confidence interval does not cover the true mean.

Given these parameters, the needed sample size can be computed as

$$n = \left(\frac{z_{\alpha/2}}{d}\right)^2 \sigma^2$$

where d, σ^2, and α are as given above and z is a standard normal variate.

Depending on the analysis objective, samples of a sufficient size may be needed to test the accuracy of specific cover types as well as the overall accuracy. Fitzpatrick-Lins discusses sample size selection in an application to land-use and land-cover mapping (3).

Representativeness of Sample. One of the important criteria for the validity of a classification accuracy estimate is that the sample on which that estimate was based must be representative of the area of interest. Use of an appropriate sampling methodology (to be discussed in the next section) is one way to help insure representativeness. In particular, stratified random sampling may be used to draw samples from within

each cover class or from separate geographic areas.

Sampling Methodology. In many cases, the test samples used for evaluation have been analyst-selected. Although this method is easy to execute, it may lead to a bias in accuracy estimation. In particular, the analyst may select fields which are easy to identify, causing the spectrally confusing fields to be omitted and resulting in an accuracy estimate which is biased upwards.

If classification accuracies are to be statistically valid, the test samples should be randomly selected. There are many random sampling methods from which to choose. Four of these, illustrated in Figure 2, are conducted as follows:

1. Simple Random Sampling. The area is divided into N blocks of a given size sampling unit (e.g., pixel). Determine the number (n) to be selected. Randomly select n of the N areas.

2. Systematic Random Sampling. Determine the sample size required and systematically locate this number of units (i.e., randomly select a starting point and sample at a fixed interval thereafter).

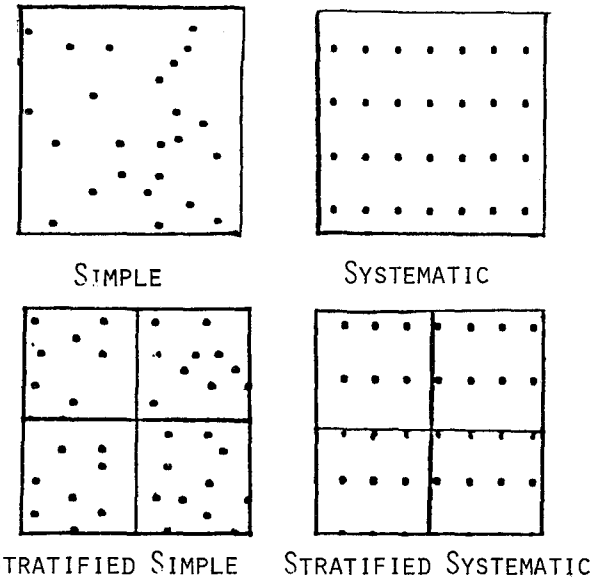

Figure 2. Schematic diagram of sample selection using four types of random sampling.

3. Stratified Simple Random Sampling. Stratify the area of interest and divide each stratum into N blocks of a given size. Determine the number of samples (n) and allocate them to the strata according to the stratification variable. Randomly select n of the N blocks.

4. Stratified Systematic Random Sampling Stratify the area of interest. Determine the number of samples (n) and allocate them to the strata according to the stratification variable. Within each stratum, systematically locate the samples by randomly selecting a starting point and sampling at a fixed interval thereafter.

Simple random sampling is most easily understood by the public, but is somewhat less convenient and less precise than some of the other sampling methods.

It is easier to draw a sample and execute the sampling procedure without errors using systematic sampling. A systematic sample is intuitively more precise than simple random sampling and is sometimes considerably more precise than stratified random sampling because the sample is spread evenly across the population.

Stratified random sampling has several advantages (4). Administrative convenience may result by dividing the work load by stratum among several individuals or field offices. This may be particularly advantageous for conducting the time-consuming task of field checking to identify test data. Stratification enables estimation of each subdivision of the populstion with known precision by considering each stratum as a "population" in its own right. Finally, stratification can provide an increased precision over simple random sampling in estimates for the entire population if the strata are homogeneous.

Selection of Sample Unit Size. To draw a simple random sample, it was necessary to divide the area of interest into N blocks of some size. These blocks are then referred to as the sampling units and an entire block (sampling unit) is measured at each location of a sample.

In remote sensing data analysis, the smallest possible sampling unit size is the pixel, but larger sampling units are

also possible. These larger sampling units are an example of a cluster sample which consists of a group or cluster of elemental units - pixels in this case (4).

Homogeneous cells have been utilized in forestry applications where an area several pixels square is defined as the sampling unit, and all pixels in that cell are utilized for the sample. The requirement for homogeneity of a block reduces the potential number of units from which the sample is selected.

Another type of sampling scheme is subsampling or two-stage sampling. First, a sample of units, known as primary sampling units, is selected and then a subsample is drawn from each sampling unit. This procedure could, for example, select sections of land for the test sample and then select agricultural fields within that section in a second stage. This type of sampling scheme is easy to execute, and is particularly well-suited for facilitating ground checking or photointerpretation. If the primary sample units are large relative to the entire area, the test set may not be representative of the area of interest.

Computation of Accuracy and Confidence Intervals. A confusion matrix or error matrix (Figure 3) is typically formed using test samples to compute classification accuracy. Several measures of accuracy can then be computed: the overall performance (the total number of pixels classified correctly divided by the total number of pixels), classification accuracy for each class or cover type, and average

Class	Total No. Samples	No. Samples Classified As		
		Corn	Soybeans	Other
Corn	981	853	9	119
Soybeans	893	4	876	13
"Other"	1397	296	93	1008

Class Performance:	Corn	853/981 = 87.0
	Soybeans	876/893 = 98.1
	Other	1008/1397 = 72.2
Average Performance:		85.8
Overall Performance:		83.7

Figure 3. Example of a confusion matrix and computation of classification accuracy estimates.

performance by class (the average accuracy of each of the cover classes tabulated).

In addition to an estimate of classification accuracy, the user may want to know what kind of variance is associated with that estimate. One way to present this type of information is by computing a confidence interval for classification accuracy. The measure of accuracy P can be considered to be distributed binomially, as a pixel is either correctly or incorrectly classified. A transformed value

$$P_T = \arcsin\sqrt{P}$$

can be considered to be distributed as normal with a standard deviation

$$S_p = \sqrt{821/n}$$

where n is the number of observations used to compute P (6). Then, following the normal properties, a 95% confidence interval for P_T is given by

$$(P_T - t_{\infty,.05}\, S_p,\ P_T + t_{\infty,.05}\, S_p)$$

and a corresponding 95% confidence interval for P is then

$$\left\{\left[\sin(P_T - t_{\infty,.05}\, S_p)\right]^2,\ \left[\sin(P_T + t_{\infty,.05}\, S_p)\right]^2\right\}$$

B. ESTIMATION OF PROPORTIONS FROM CLASSIFICATION RESULTS

Once a classification of Landsat data has been carried out, the results of this analysis will be used to estimate the area or proportion of the cover types of interest. Four methods will be discussed: classify and count, bias correction, the stratified areal estimate, and regression estimation.

The classify and count method is straightforward: the proportion estimate is given as

$$\hat{P}_i = \frac{n_i}{n}$$

where n is the number of pixels classified as cover type i and n is the number of pixels in the sample. This method is direct but is biased unless errors of omission and commission cancel out.

Bias in area or proportion estimates can be removed if classification error rates are known. The error or confusion matrix discussed in the previous section provides an estimate of the classification error rates. Denoting the

error matrix by E, a bias corrected estimate can be computed as

$$P = (E^T)^{-1} \hat{P}$$

since

$$\hat{P} = E^T P$$

where P is the vector of true proportions and P is the vector of classify and count proportions (7). This technique for bias correction has had mixed results. Keys to its successful use appear to be representative test fields (to obtain a good estimate of E) and relatively high classification accuracies for all cover types of interest.

Another method for computing an unbiased proportion estimate is the stratified areal estimate (SAE) used in LACIE (10). All pixels classified into class i are considered to form stratum i. Test samples are used to find

$$\alpha_{ij} = \frac{n_{ij}}{n_j}$$

where n_{ij} is the number of samples in stratum j which belong to class i and n_j is the number of test samples in stratum j.

An estimate of the proportion of cover type i is

$$\hat{P}_i = \sum_j \alpha_{ij} \frac{n_j}{n}$$

where n_j is the number of pixels classified as class j and n is the total number of pixels in the area. Using conditional probability notation, this can be represented as

$$\hat{P}_i = \sum_j \text{Prob}\,(C_i | \hat{C}_j)\ \text{Prob}\,(\hat{C}_j).$$

The SAE is an unbiased estimation method and is relatively easy to compute. However, in the selection of test samples for use with this method, care must be taken as the method assumes test samples are proportionally allocated to classes.

A fourth method for area or proportion estimation is regression estimation (4,11). This method has been used by the USDA/ESS with positive results in several states. The regression combines the use of ground data and Landsat classifications to produce estimates with improved precision over the use of ground data only and reduced bias over the use of Landsat data only. Disadvantages are that it requires a large area to be classified and is liable to bias if the samples are small relative to the individual strata.

C. COMPARISON OF AREA OR PROPORTION ESTIMATES WITH A REFERENCE STANDARD

In evaluation of results where the analysis objective has been area or proportion estimation, the computation of the classification accuracy is only one step in the evaluation process. In particular, the quality of the area or proportion estimate itself should be evaluated by comparison with some form of reference data. Two type of comparisons will be discussed: (1) a correlation between Landsat-derived estimates and the reference data and (2) a test of hypothesis comparing the Landsat estimates and reference data at an appropriate significance level (alpha-level).

Correlation Between Landsat Estimates and Reference Data. If independent area estimates have been made for several areas, then a correlation between the Landsat-derived estimates and a reference standard can be computed (Figure 4). In addition to a high correlation which indicates a strong relationship between the two quantities, a one-to-one relationship is desirable; i.e., if the points fall about a 45 degree line, this indicates a lack of bias in the estimation procedure.

Test of Hypothesis. For more than a qualitative evaluation, a test of hypothesis can be conducted to compare the classification estimates with the reference data. Two types of tests are available: parametric and nonparametric

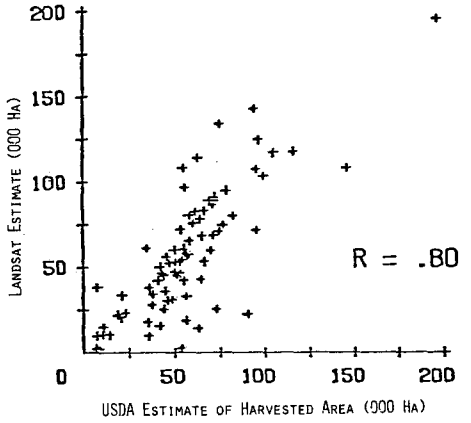

Figure 4. Correlation between USDA estimate and Landsat classification estimate for the area of wheat in Kansas (7). Each point represents a county estimate.

tests. In parametric tests (such as the t-test and analysis of variance), the normal distribution and equality of variances are assumed (8). These tests are reasonably robust to departures from normality, but care in the interpretation of test results should be taken unless the assumptions are strictly satisfied.

If these assumptions of normality and homogeneity are not satisfied, then a nonparametric statistical test should be utilized (9). This family of tests does not make any assumptions about the form of the distribution. However, since they are generally not as powerful as parametric tests, a nonparametric test should be used only when the parametric assumptions cannot be met.

D. COMPARISON OF AREA OR PROPORTION ESTIMATES WITH RESULTS FROM ANOTHER ANALYSIS

A researcher might want to compare analysis or estimation methods, asking the questions:

1. "Are the methods different in accuracy or in resulting estimates?"

2. "Which methods are significantly different and which is the best method or group of methods to use?"

To address the first question, analysis of variance (ANOVA) is an appropriate analytical tool. ANOVA tests to see if a factor (e.g., analysis method, classifier, etc.) has a significant effect on a dependent variable (e.g., classification accuracy or resulting estimates). ANOVA is a parametric test and, as such, assumes normality and homogeneity of variance for the dependent variable. Percent data (such as overall percent correct) can often be made to satisfy these assumptions by using a transformation (6,8).

The results of an analysis of variance will indicate whether or not methods have a significant effect on classification accuracy. The ANOVA does not, however, tell which method is best; to address this question, a multiple range test should be used. Many multiple range tests are available such as the Newman-Keuls, Duncan, and Tukey procedures. The multiple range test is performed on factors which ANOVA has found to be significant. It determines, at a specified alpha level, which methods or levels of the factor are significantly different from one another.

IV. SUMMARY

In the past several years, there has been an increasing awareness in the remote sensing community of the need for statistical results evaluation. Conference sessions, workshops, and journal papers have been devoted to this topic. In this paper, I have tried to present some of the considerations for evaluating area estimates.

In summary, I would like to encourage each individual and organization to continue to stress results evaluation. This is not a field which should cause great apprehension: an introductory statistics textbook (not requiring calculus) and a textbook on sampling theory should enable most remote sensing scientists to be well on their way toward the evaluation and documentation of the significance of their analysis results.

V. REFERENCES

1. MacDonald, R.B., and F.G. Hall. 1980. Global Crop Forecasting. Science 208:670-679.

2. Houston, A.G., A.H. Feiveson, R.S. Chhikara, and E.M. Hsu. 1979. Accuracy Assessment: The Statistical Approach to Performance Evaluation in LACIE. Proc, The LACIE Symp., Houston, Texas, October 23-26, 1978, pp. 115-130. JSC-16015.

3. Fitzpatrick-Lins, Katherine. 1981. Comparison of Sampling Procedures and Data Analysis for a Land-Use and Land-Cover Map. Photog. Engin. 47:343-351.

4. Cochran, William G. 1963. Sampling Techniques. John Wiley & Sons, Inc., New York.

5. Bizzell, R.M., F.G. Hall, A.H. Feiveson, M.E. Bauer, B.J. Davis, W.A. Malila, and D.P. Rice. 1975. Results from the Crop Identification Technology Assessment for Remote Sensing (CITARS) Project. Proc., Tenth Int'l Symp. on Remote Sensing of Environment, Ann Arbor, Michigan.

6. Bartlett, M.S. 1947. The Use of Transformations. Biometrics 3:39-52.

7. Bauer, Marvin E., Marilyn M. Hixson, Barbara J. Davis, and Jeanne B. Etheridge. 1978. Area Estimation of Crops by Digital Analysis of Landsat Data. Photog. Engin. 44:1033-1043.

8. Anderson, Virgil L. and Robert A. McLean. 1974. Design of Experiments: A Realistic Approach. Marcel Dekker, Inc., New York.

9. Hollander, Myles, and Douglas A. Wolfe. 1973. Nonparametric Statistical Methods. John Wiley & Sons, Inc., New York.

10. Heydorn, R.P., R.M. Bizzell, J.A. Quirein, K.M. Abotteen, and C.A. Sumner. 1979. Classification and Mensuration of LACIE Segments. Proc., The LACIE Symp., Houston, Texas, October 23-26, 1978, pp. 73-86. JSC-16015.

11. Hanuschak, George, Richard Sigman, Michael Craig, Martin Ozga, Raymond Luebbe, Paul Cook, David Kleweno, and Charles Miller. 1979. Crop-Area Estimates from Landsat; Transition from Research and Development to Timely Results. Proc., Machine Processing of Remote Sensed Data Symp., West Lafayette, Indiana, pp. 86-96.

AUTHOR BIOGRAPHICAL DATA

Marilyn M. Hixson is Senior Research Statistician in Crop Inventory Research at LARS. She holds a B.S. in mathematics from Miami University and an M.S. in mathematical statistics from Purdue University. Ms. Hixson has had a major role in the design, Landsat data classifications, and statistical analysis of results in several Landsat investigations concerning training, classification, and area estimation procedures for crop inventory, including both segment and full-frame sampling approaches. Her work on field research projects has involved experiment design, data analysis, and statistical consulting. She is a member of the American Statistical Association and the American Society of Photogrammetry.