Russell G. Congalton
Richard G. Oderwald
Roy A. Mead*
School of Forestry and Wildlife Resources
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061

Assessing Landsat Classification Accuracy Using Discrete Multivariate Analysis Statistical Techniques

These techniques allow the Landsat data user to quantitatively compare the different aspects of image processing and to determine which perform better under varied conditions.

Introduction

The need for techniques to assess the accuracy of Landsat derived information cannot be understated. Without methods for measuring and comparing the accuracy attained using various classification schemes, improvement of these schemes is impossible. The objective of this paper is to inform the users of remotely sensed data of some relatively new and unknown techniques for assessing Landsat classification accuracy. It is not within the scope of this paper to present all the theoretical and practical details of the statistical analysis involved here. However, it is hoped that even those users with little to no statistical background will be able to see the usefulness of these techniques from this presentation.

The method of accuracy assessment described here uses discrete multivariate analysis techniques. These techniques are appropriate because they are designed for the analysis of discrete data. Classification data are discrete because the data either fall into a particular land-cover category or they do not. For example, a pixel can be classified as pine, hardwood, or water but not as half pine and half water. Most previous accuracy assessment techniques have used parametric statistical techniques which assume continuous data and normal distributions.

Abstract: *Discrete multivariate analysis techniques have been used to evaluate the accuracy of land-cover classifications from Landsat digital imagery. Error matrices or contingency tables were taken from the literature and then analyzed using three techniques. The first technique permitted direct comparison of corresponding cell values in different matrices by "normalizing" each matrix through a process called "iterative proportional fitting." The second technique provided a method of testing for significant differences between error matrices which vary by only a single variable. The third technique allowed for multivariable comparisons between matrices to be made and is the most powerful of the techniques. It was concluded that these techniques could help researchers better evaluate variables or factors affecting classification accuracy.*

The most common way to represent the accuracy of a Landsat classification is in the form of an error matrix or contingency table (e.g., Card, 1982; Mead and Meyer, 1977; Hoffer, 1975). An error matrix is a square array of numbers set out in rows and columns which express the number of pixels assigned as a particular land-cover type relative to the actual land cover as verified in the field or from interpreted aerial photographs. The columns usually represent the reference data (i.e., assumed correct) and

* Presently employed by Technicolor Government Services, Inc., Denver, Colorado.

Photogrammetric Engineering and Remote Sensing,
Vol. 49, No. 12, December 1983, pp. 1671-1678.

0099-1112/83/4912-1671$02.25/0

the rows indicate the computer assigned land-cover category (i.e., Landsat data).

This form of expressing accuracy as an error matrix is an effective way to evaluate both errors of inclusion (commission errors) and errors of exclusion (omission errors) present in the classification. Also, the error matrix allows the analyst to determine the performance for individual categories as well as for the overall classification (Hoffer and Fleming, 1978). In the ideal situation, all the non-major diagonal elements of the error matrix would be zero, indicating that no pixel had been misclassified.

Once the error matrix has been generated, a very simple procedure can be used to determine the overall accuracy. Because the values on the major diagonal represent those pixels that have been correctly classified, these values are summed up and divided by the total number of pixels classified. This number is then the overall performance accuracy of an error matrix, and is the most common use of the error matrix in accuracy assessment.

Until recently, this measure of overall performance accuracy was the extent of most accuracy assessments. However, additional statistical techniques are now being used to further assess classification accuracy. These methods can be divided into two groups: analysis of variance and discrete multivariate analysis (often called contingency table analysis).

Analysis of variance makes use of only the diagonal elements in the error matrix. Also, the technique requires that the data be normally distributed. As previously mentioned, classification data are discrete and multinomially distributed (each category is binomially distributed). The diagonal elements of the error matrix can be converted to a normal distribution using various transformations (Snedecor and Cochran, 1976). However, another assumption of analysis of variance is that the categories in the error matrix are independent. This assumption is often not met in remotely sensed data because of the confusion between categories. For additional details and examples of this technique, see Rosenfield (1982).

Discrete multivariate analysis, on the other hand, does not assume that the categories are independent nor does it require any transformation of the data. Instead, these techniques are designed specifically to deal with categorical data. Discrete multivariate analysis also uses the entire error matrix and not just the diagonal elements. As suggested by Card (1982), "contingency table analysis is the most natural framework for accuracy assessment, both for the convenient display of empirical results and for the ease of statistical analysis."

Discrete Multivariate Analysis Techniques

Three different methods of comparing error matrices using discrete multivariate analysis techniques were evaluated in this study. The first method allows for direct comparison of error matrices through a process called normalization. The second method computes a measure of agreement between error matrices which can be used to test if the matrices are significantly different. The third method provides for the simultaneous examination of all factors affecting the classification.

The first comparison procedure (Bishop *et al.*, 1975) allows corresponding cell values in different error matrices to be directly compared. This comparison is made possible by a standardizing process called normalization. Normalization of an error matrix is performed by a procedure called "iterative proportional fitting." The rows and columns of a matrix are successively balanced until each row and each column adds up to a given value (say 1.0). This process forces each cell value to be influenced by all the other cells values in its corresponding row and column. Each cell is then a combination of ground truth and computer classification and is representative of both omission and commission errors for that land-cover category.

Prior to the normalization procedure, comparison of corresponding cell values in different matrices was only possible if the matrices had the same sample size. Even then, the cell value may have been misleading because errors of omission and commission were ignored. However, due to the normalization procedure, the corresponding cell values of two or more error matrices can now be compared directly without regard for differences in sample size and including omission and commission errors. Although there is no test for significance between corresponding cell values, direct comparison can provide a relative measure of which is better because all columns and rows in the matrices are required to sum to a certain marginal.

The normalization procedure converges to a unique set of maximum likelihood estimates and as such is the most appropriate algorithm to use in this case (Fienberg, 1970). However, an assumption made by this procedure is that all cells are of equal weight or importance. This assumption is not always valid in remotely sensed data. However, it is possible to modify the fitting procedure to deal with categories of unequal importance (Bishop *et al.*, 1975). In either case, this method does provide a way of eliminating the effect of sample size while incorporating omission and commission errors into the accuracy measurement.

The second method of comparison examined here is a procedure that tests if the overall agreement in two separate error matrices is significantly different. A measure of overall agreement is computed for each matrix based on the difference between the actual agreement of the classification (i.e., agreement between computer classification and reference data as indicated by the diagonal elements) and the chance agreement which is indicated by the product of the row and column marginals. This measure of agreement, called KHAT (i.e., $\hat{K}$), is calculated by

$$\hat{K} = \frac{\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r} (x_{i+} * x_{+i})}{N^2 - \sum_{i=1}^{r} (x_{i+} * x_{+i})}$$

where r is the number of rows in the matrix, x_{ii} is the number of observations in row i and column i (i.e., the i^{th} diagonal element), x_{i+} and x_{+i} are the marginal totals of row i and column i, respectively, and N is the total number of observations (Bishop *et al.*, 1975). Notice that the numerator of this equation is similar to the observed minus the expected calculation performed in a chi-square analysis.

A KHAT value is computed for each matrix and is a measure of how well the classification agrees with the reference data (i.e., a measure of overall accuracy). Confidence intervals can be calculated for KHAT using the approximate large sample variance (Bishop *et al.*, 1975, p. 396)

$$\hat{\sigma}(\hat{K}) = \frac{1}{N} \frac{\theta_1(1 - \theta_1)}{(1 - \theta_1)^2} + \frac{2(1 - \theta_1)(2\theta_1\theta_2 - \theta_3)}{(1 - \theta_2)^3} + \frac{(1 - \theta_1)^2(\theta_4 - 4\theta_2)^2}{(1 - \theta_2)^4}$$

where

$$\theta_1 = \sum_{i=1} \frac{x_{ii}}{N} \qquad \theta_3 = \sum_{i=1} \frac{x_{ii}}{N}\left(\frac{x_{i+}}{N} + \frac{x_{+i}}{N}\right)$$

$$\theta_2 = \sum_{i=1} \frac{x_{i+} * x_{+i}}{N^2} \qquad \theta_4 = \sum_{\substack{i=1 \\ j=1}} \frac{x_{ij}}{N}\left(\frac{x_{j+}}{N} + \frac{x_{+i}}{N}\right)^2.$$

A test for significance of KHAT can be performed for each matrix separately to determine if the agreement between the classification and the reference data is significantly greater than zero. In other words, a test can be performed to see if the classification is significantly better than a random assignment of land-cover categories to pixels. More importantly, a pairwise test of significance can be performed between two independent KHAT's using the normal curve deviate to determine if the two error matrices are significantly different (Cohen, 1960). The test statistic for significant difference in large samples is given by

$$Z \sim \frac{\hat{K}_1 - \hat{K}_2}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}}. \qquad (1)$$

The confidence intervals and significance tests are based on the asymptotic normality of the KHAT ($\hat{K}$) statistic.

The above test between two independent KHAT's allows any two error matrices to be compared in order to determine if they are significantly different. In other words, error matrices generated from several classification algorithms can now be compared, two at a time, to determine which classifications are significantly better than the rest. Researchers can also use this procedure to test the effects of individual factors on the accuracy of the classification. However, this procedure would be limited in that only one factor in the classification may vary at a time. For example, in order to determine which date of imagery yields the best results, all other factors (i.e., algorithm, analyst, Landsat scene, etc.) must be held constant. Actually, this condition is fairly common in accuracy assessments; therefore the procedure can be quite useful.

The third method of comparison allows one to simultaneously analyze more than a single factor affecting the classification accuracy. The log-linear model approach as described by Fienberg (1980) and Bishop *et al.* (1975) is a method of comparison by which many variables (factors) affecting the accuracy and their interactions can be tested together to determine which are necessary (i.e., significant) in fully explaining the classification accuracy.

In this method, the simplest model (combination of variables and their interactions) that provides a good fit to the data (error matrices) is chosen using a model selection procedure. This procedure, which is similar to model selection procedures used in regression (i.e., forward selection, etc.), allows the user to systematically search all possible combinations of variables and their interactions and choose the simplest combination that provides a good fit to the data. First, all uniform order log-linear models (i.e., models with all possible n-way interactions, where $n = 1$ to the number of variables) are examined and the simplest good fit model is chosen. Each interaction of the chosen model is then tested for significance. If the interaction is not significant, it is dropped from the model. This process continues for each interaction until a model is found in which all the factors and interactions are significant. A more detailed description of this stepwise model selection procedure can be found in Section 5.3 of Fienberg (1980). The criteria for determining the significance of a model are based on the Likelihood Ratio, G^2, and the corresponding degrees of freedom for the model.

This procedure uses a method of successive approximations (i.e., "iterative proportional fitting") which converges to the maximum likelihood estimates of the minimum sufficient statistics as defined by the model. In other words, the "iterative proportional fitting" procedure attempts to fit the model of interest to the data. This procedure is very tedious and time-consuming and is almost always done on the computer. The Likelihood Ratio, G^2, is then used as a measure of "goodness of fit" of the model to the data. The Likelihood Ratio statistic, G^2, (Equation 2) is used in place of the Pearson chi-square statistic, χ^2, (Equation 3) because G^2 can be partitioned, as in the model selection procedure, and still retain an approximate chi-square distribution. Therefore, the critical value for testing if the

model of interest is a good fit can be obtained from a chi-square table with the appropriate degrees of freedom (Fienberg, 1980; Bishop *et al.*, 1975).

$$G^2 = 2\Sigma \text{ (observed) } \log\left(\frac{\text{observed}}{\text{expected}}\right) \quad (2)$$

$$\chi^2 = \Sigma \frac{(\text{observed-expected})^2}{\text{expected}} \quad (3)$$

Therefore, the log-linear model approach allows for analysis of multi-way error matrices with many factors. For example, error matrices generated using different dates, different algorithms, and different analysts all of the same scene of imagery can be put together and the factors necessary to explain the classification accuracy determined. A possible practical result would be that the date of the imagery was insignificant and therefore the date of the imagery could be selected with other objectives in mind.

It should be realized here that performing any of these three comparison methods by hand would be very tedious. Computer programs have been written to implement all three techniques (Congalton *et al.*, 1981; Congalton *et al.*, 1982). However, due to space limitations, these programs will not be presented in this paper.

Example Analyses

The data used to test the individual effects of different classification algorithms on Landsat classification accuracy were part of a study done by Hoffer (1975) on mountainous terrain in southwest and central Colorado. Each algorithm was used to classify the image into one of four broad land-cover categories: conifer (C), deciduous (D), agriculture (A), and water (W). The four classification algorithms used were a non-supervised 20-cluster algorithm, a non-supervised 10-cluster algorithm, a modified supervised algorithm, and a modified clustering algorithm. The original and the normalized error matrices for each of the four algorithms are given in Tables 1 to 4. Table 5 presents a comparison of overall performance and normalized performance accuracies for the four algorithms. Note that both accuracies yield the same ranking (best to worst) except for the non-supervised 20-cluster algorithm. This discrepancy is due to the weighting of the major diagonal by the other cell values in its row and column.

As previously mentioned, normalization allows direct comparison of corresponding cell values in each of the four error matrices. For example, Table 6 contains the percent correct and normalized values for the deciduous category in each matrix. In the percent correct calculation only errors of omission are accounted for. However, the normalized value considers both omission and commission errors as well as negating the effect of sample size.

Table 1. The Original and Normalized Error Matrices for a Non-supervised 10 Cluster Classification of the Ludwig Mountain Area

Reference Data

Landsat Data	C	D	A	W	
C	317	23	0	0	
D	61	120	0	0	N = 659
A	2	4	60	0	$\frac{\Sigma_{++}}{N} = \frac{505}{659} = .7663$
W	35	29	0	8	
					$\Sigma_{++} = 505$

Reference Data

Landsat Data	C	D	A	W	
C	0.7767	0.1340	0.0171	0.0718	
D	0.1623	0.7415	0.0184	0.0775	N = 4
A	0.0028	0.0119	0.9531	0.0332	$\frac{\Sigma_{++}}{N} = \frac{3.2888}{4} = .8222$
W	0.0581	0.1126	0.0114	0.8175	
					$\Sigma_{++} = 3.2888$

The result of considering both omission and commission errors in the accuracy figure is clearly demonstrated by comparing the values of the two non-supervised algorithms for the deciduous category. Notice that the 10-cluster algorithm classified 120 out of 176 deciduous pixels correctly for a percent correct value of 68 percent, while the 20-cluster algorithm classified only 72 out of 176 deciduous pixels correctly for a percent correct of only 41 percent. However, notice that the normalized value for the 10-cluster algorithm is 0.7415 while the normalized value for the 20-cluster algorithm is 0.7817. This apparent discrepancy in results is due to the large commission error in the 10-cluster algorithm error matrix (see Table1). The practical application of this result is that perhaps the 10-cluster algorithm is not that much better than the 20-cluster algorithm at classifying the deciduous category, despite what is indicated by the percent correct value.

The data supplied by Hoffer (1975) were also used

TABLE 2. THE ORIGINAL AND NORMALIZED ERROR MATRICES FOR A NON-SUPERVISED 20 CLUSTER CLASSIFICATION OF THE LUDWIG MOUNTAIN AREA

Reference Data

Landsat Data	C	D	A	W	
C	377	79	0	0	
D	2	72	0	0	N = 659
A	33	5	60	0	$\frac{\Sigma_{++}}{N} = \frac{517}{659} = 0.7845$
W	3	20	0	8	
					$\Sigma_{++} = 517$

Reference Data

Landsat Data	C	D	A	W	
C	0.8606	0.1155	0.0080	0.0157	
D	0.0423	0.7817	0.0593	0.1164	N = 4
A	0.0716	0.0075	0.9071	0.0147	$\frac{\Sigma_{++}}{N} = \frac{3.4026}{4} = .8506$
W	0.0255	0.0953	0.0256	0.8532	
					$\Sigma_{++} = 3.4026$

TABLE 3. THE ORIGINAL AND NORMALIZED ERROR MATRICES FOR A MODIFIED SUPERVISED CLASSIFICATION OF THE LUDWIG MOUNTAIN AREA

Reference Data

Landsat Data	C	D	A	W	
C	305	64	0	0	
D	90	94	0	0	N = 646
A	18	13	60	0	$\Sigma_{++} = \frac{461}{646} = 0.7136$
W	0	0	0	2	
					$\Sigma_{++} = 461$

Reference Data

Landsat Data	C	D	A	W	
C	0.6671	0.2941	0.0103	0.0276	
D	0.2963	0.6461	0.0154	0.0414	N = 4
A	0.0294	0.0448	0.9064	0.0201	$\frac{\Sigma_{++}}{N} = \frac{3.1305}{4} = 0.6261$
W	0.0072	0.0150	0.0679	0.9109	
					$\Sigma_{++} = 3.1305$

to test for significant differences between error matrices. The error matrices generated from the four classification algorithms can be tested to see which are significantly different. The KHAT statistic can then be used as an accuracy measurement (i.e., measure of agreement) to determine which of the significantly different matrices and hence algorithms are best. Table 7 shows the results of the pairwise significance tests. This table also contains the KHAT value and its associated variance for each matrix. Notice that the test between the non-supervised 10-cluster algorithm and the non-supervised 20-cluster algorithm is not significant. This result indicates that there is no justification for spending the extra time to use the 20-cluster approach because the 10-cluster approach works just as well. All other pairwise combinations of error matrices are significantly different. Therefore, based on KHAT values, the modified clustering algorithm is the best, while the modified supervised algorithm is the worst at classifying this image. These results agree with the overall performance accuracy values given in Table 5. They also agree with normalized performance accuracy values given in Table 5 except for some confusion between the modified clustering and 20-cluster non-supervised approaches. In this case the normalized values for the two approaches differ by only 0.003. However, this confusion does confirm that the use of omission and commission errors in the accuracy measurement (normalization) can alter the results completely.

The data used to test the multifactor effect of different classification algorithms and enhancement techniques on classification accuracy were supplied by Gregg *et al.* (1979). These data were collected as part of an operational study of Landsat imagery for inventory purposes in the State of Washington. In this example, two classification algorithms, two enhancement techniques, ten reference data categories, and ten Landsat data categories were studied

TABLE 4. THE ORIGINAL AND NORMALIZED ERROR MATRICES FOR A MODIFIED CLUSTERING CLASSIFICATION OF THE LUDWIG MOUNTAIN AREA

Reference Data

Landsat Data	C	D	A	W	
C	379	50	0	5	
D	19	101	0	0	$N = 632$
A	7	7	60	0	$\frac{\Sigma_{++}}{N} = \frac{543}{632} = 0.8592$
W	1	0	0	3	
					$\Sigma_{++} = 543$

Reference Data

Landsat Data	C	D	A	W	
C	0.7860	0.1222	0.0040	0.0874	
D	0.1355	0.8240	0.0133	0.0267	$N = 4$
A	0.0299	0.0349	0.9209	0.0153	$\Sigma_{++} = \frac{3.4015}{4} = 0.8503$
W	0.0486	0.0189	0.0619	0.8706	
					$\Sigma_{++} = 3.4015$

resulting in a four-way table of dimension 2 by 2 by 10 by 10. Unfortunately, due to the size of this four-way table, it cannot be printed here. However, the original data can be found in the paper cited above.

As previously described, a model selection procedure was used to determine the simplest good fit model to the data. Table 8 contains the uniform order log-linear models for these data. Notice that, because this is a four-way table, the uniform order models consist of the models with all three-way, all two-way, and all one-way interactions. The uniform order model of all four-way interactions (i.e., [1 2 3 4]) is the complete or saturated model and will always fit the data. However, the object here is to find the simplest good fit model.

In this example, classification algorithm (denoted variable [1]), enhancement technique (denoted variable [2]), and the reference data (denoted variable [3]) are called the explanatory variables while the Landsat data (denoted variable [4]) are called the response variable. This terminology results because the first three variables are being used to try to explain the response (i.e., Landsat classification). The interaction terms in the model are represented as combinations of these variables enclosed in brackets (e.g., [1 2] is the interaction between algorithm and enhancement).

Table 8 shows that the two-way interaction uniform order model is the simplest good fit model as determined by the Likelihood Ratio, G^2. Therefore, this model will be used as the first step in the model selection procedure. Notice that this model contains six two-way interaction terms. The object then is to systematically remove all other non-significant factors or their interactions. Table 9 shows the steps of

TABLE 5. A COMPARISON OF THE OVERALL PERFORMANCE AND THE NORMALIZED PERFORMANCE ACCURACY FOR THE FOUR ALGORITHMS

Algorithm	Overall Performance Accuracy	Normalized Performance Accuracy
modified clustering	0.8592	0.8503
nonsupervised 20 clusters	0.7845	0.8506
nonsupervised 10 clusters	0.7663	0.8222
modified supervised	0.7136	0.6261

TABLE 6. A COMPARISON OF THE DECIDUOUS CLASSIFICATION FOR THE FOUR CLASSIFICATION ALGORITHMS.

Algorithm	Number of Correctly Classified Pixels	Number of Pixels in the Deciduous Category	% Correct	Normalized Value
nonsupervised 10 clusters	120	176	0.6818	0.7415
nonsupervised 20 clusters	72	176	0.4090	0.7817
modified supervised	94	171	0.5497	0.6461
modified clustering	101	158	0.6392	0.8240

TABLE 7. THE RESULTS FOR THE TEST OF AGREEMENT BETWEEN ERROR MATRICES FOR THE FOUR CLASSIFICATION ALGORITHMS

Pairwise Comparison	Z Statistic[1]	Result 95%	Result 90%
(10 N-S) & (20 N-S)	0.475	NS[2]	NS
(10 N-S) & (MS)	3.009	S	S
(10 N-S) & (MC)	−2.936	S	S
(20 N-S) & (MS)	2.434	S	S
(20 N-S) & (MC)	−3.281	S	S
(MS) & (MC)	−5.624	S	S

Error Matrix	KHAT Statistic	Variance
10 cluster non-supervised (10 N-S)	0.605	0.00073735
20 cluster non-supervised (20 N-S)	0.586	0.00087456
Modified Supervised (MS)	0.476	0.00109972
Modified Clustering (MC)	0.718	0.00076218

[1] Use equation (1) to calculate Z statistic.
[2] NS = non-significant result; S = significant result.

this process. The next step in the model selection process then is to eliminate a two-way interaction term from the two-way uniform order model. Six new models result, each with a different combination of the five remaining two-way interaction terms. These six new models are tested for "goodness of fit" based on the Likelihood Ratio, G^2, and the appropriate degrees of freedom (df), and model B is chosen to be the simplest best fit model. The missing two-way interaction term is tested for significance by comparing the fit of the two-way uniform order model, labeled A, (see Table 8) with model B. The test is possible because partititoning the Likelihood Ratio still results in a chi-square distribution. Because the test is not significant, the [2 3] interaction term is dropped from the model (see Table 9).

TABLE 8. THE UNIFORM ORDER MODELS FOR THE FOUR-WAY TABLE COMPARING ENHANCEMENT TECHNIQUES AND CLASSIFICATION ALGORITHMS

Model	G^2	df	Result
[1][2][3][4]	10888.87281	352	poor fit
[12][13][14][23][24][34][A]	145.86428	234	good fit
[123][124][134][234]	20.90917	54	good fit

TABLE 9. THE MODEL SELECTION PROCESS FOR THE FOUR-WAY TABLE COMPARING ENHANCEMENT TECHNIQUES AND CLASSIFICATION ALGORITHMS

Model	G^2	df	Result
[12][13][14][23][24]	10732.65712	315	poor fit
[12][13][14][23][34]	230.22104	243	good fit
[12][13][14][24][34][B]	147.83246	243	good fit
[12][13][23][24][34]	227.51772	243	good fit
[12][14][23][24][34]	156.53131	243	good fit
[13][14][23][24][34]	146.04061	235	good fit

model B best and good fit not significant
$G^2(B) - G^2(A) = 1.96818 \sim \chi^2_{9df}$ so drop [23]

Model	G^2	df	Result
[12][13][14][24]	10733.8278	324	poor fit
[12][13][14][34]	231.3918	252	good fit
[12][13][24][34]	229.4802	252	good fit
[12][14][24][34]	158.4941	252	good fit
[13][14][24][34][C]	148.0034	252	good fit

model C best and good fit not significant
$G^2(C) - G^2(B) = 0.17094 \sim \chi^2_{1df}$ so drop [12]

Model	G^2	df	Result
[13][14][24]	10733.99876	325	poor fit
[13][14][34][2]	231.43485	253	good fit
[13][24][34]	229.52108	253	good fit
[14][24][34][D]	158.66500	253	good fit

model D best and good fit not significant
$G^2(D) - G^2(C) = 10.66162 \sim \chi^2_{1df}$ so drop [13]

Model	G^2	df	Result
[14][24][3]	10734.29202	334	poor fit
[14][34][2]	242.09646	262	good fit
[24][34][1][E]	229.81433	262	good fit

model E best and good fit choose D
$G^2(E) - G^2(D) = 71.14933 \sim \chi^2_{9df}$ significant so can't drop [14]

This same process is repeated, yielding model C as the simplest best fit model. Also, the test shows the [1 2] interaction term to be not significant; therefore, it is dropped from the model. Again the process is repeated, leading to model D and the

elimination of the [1 3] interaction. Note that one of the possible models tested here contained a one-way interaction term. The process is repeated one last time, resulting in model E as the simplest best fit model. However, the test between models D and E was significant; therefore, the [1 4] interaction cannot be dropped from the model without losing some information about the data. Therefore, model D ([1 4] [2 4] [3 4]) is selected as the simplest best fit model to the data.

Model D indicates that there are no three-way interactions necessary to explain the data. Instead, there is a combined effect due to each explanatory variable (i.e., algorithm, enhancement, and reference data) separately with the response variable. In other words, for this example each factor is significant in the performance of classifying the image and, therefore, none can be eliminated.

Conclusions

The three techniques described here should be very helpful in comparing and assessing Landsat classification accuracy data that are in the form of error matrices. These techniques allow the Landsat data user to quantitatively compare the different aspects of image processing and determine which perform better under varied conditions. Wide use of these quantitative methods could lead to greater improvement in our application of Landsat imagery.

Acknowledgments

This work was supported by the Nationwide Forestry Applications Program, Renewable Resources Inventory Project Cooperative Agreement Number 13-1134. The authors would like to thank the anonymous reviewers of this manuscript for their comments and suggestions.

References

Bishop, Y, S. Fienberg, and P. Holland, 1975. *Discrete Multivariate Analysis—Theory and Practice.* MIT Press: Cambridge, Massachusetts. 575 p.

Card, Don H., 1982. Using known map category marginal frequencies to improve estimates of thematic map accuracy. *Photogrammetric Engineering and Remote Sensing*, Vol. 48, No. 3, pp. 431–439.

Cohen, Jacob, 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, Vol. XX, No. 1, pp. 37–40.

Congalton, R., R. Mead, R. Oderwald, and J. Heinen, 1981. *Analysis of forest classification accuracy.* Remote Sensing Research Report 81-1. Virginia Polytechnic Institute and State University. Nationwide Forestry Applications Program Cooperative Research Report No. 13-1134. 85 p.

Congalton, R., R. Oderwald, and R. Mead, 1982. *Accuracy of remotely sensed data: sampling and analysis procedures.* Remote Sensing Research Report 82-1. Virginia Polytechnic Institute and State University. Nationwide Forestry Applications Program Cooperative Research Report No. 13-1134. 82 p.

Fienberg, S., 1970. An iterative procedure for estimation in contingency tables. *Annals of Mathematical Statistics*, Vol. 41, No. 3, pp. 907–917.

———, 1980. *The Analysis of Cross-Classified Categorical Data.* MIT Press: Cambridge, Massachusetts. 198 p.

Gregg, T., E. Barthmaier, R. Aulds, and R. Scott, 1979. *Landsat operational inventory study.* Division of Technical Services. State of Washington. Department of Natural Resources. Olympia, Washington. 28 p.

Hoffer, R., 1975. *Natural resource mapping in mountainous terrain by computer analysis of ERTS-1 satellite data.* Purdue University. LARS Research Bulletin 919. 124 p.

Hoffer, R., and M. Fleming, 1978. *Mapping vegetative cover by computer aided analysis of satellite data.* Purdue University. LARS Technical Report 011178. 10 p.

Mead, R., and M. Meyer, 1977. Landsat digital data application to forest vegatation and land use classification in Minnesota. In *Machine Processing of Remotely Sensed Data, Proceedings.* Purdue Unnversity. pp. 270–280.

Rosenfield, George H., 1982. *Analyzing thematic maps and mapping for accuracy.* United States Geological Survey Open File Report 82–239. 16 p.

Snedecor, G., and W. Cochran, 1976. *Statistical Methods.* Sixth Edition. The Iowa State Press, Ames, Iowa. 622 p.

(Received 11 May 1982; revised and accepted 14 July 1983)

CALL FOR PAPERS

Ninth Canadian Symposium on Remote Sensing

Memorial University of Newfoundland, St. John's, Newfoundland
13-17 August 1984

The theme chosen for this symposium—sponsored by the Canadian Remote Sensing Society—is "Remote Sensing for the Development and Management of Frontier Areas," with emphasis on oceans, the northland, and wilderness regions. The conference will consist of plenary, technical, and poster sessions.

The Technical Program Committee invites authors to submit a 600-word abstract of papers proposed for presentation at the symposium, no later than 29 *February 1984*, to the following address:

Dr. Denes Bajzak
Faculty of Engineering and Applied Science
Memorial University of Newfoundland
St. John's, Newfoundland A1B 3X5, Canada