

**INTEGRATED TRAINING AREA MANAGEMENT  
ITAM Learning Module  
LCTA Scenario**

**Accuracy Assessment of the Discrete Classification of  
Remotely-Sensed Digital Data**

---

## **Recommended Reading**

ITAM Technical Reference Manual:

Chapter 8: *Structured Query Language (SQL)*

Chapter 11: *Data Analysis and Interpretation*

Senseman, G.M., Bagley, C.F., and S.A. Tweddale. 1995. *Accuracy Assessment of the discrete Classification of Remotely-Sensed Digital Data for Landcover Mapping*. USACERL Technical Report EN-95/04, April 1995.

## **Problem Statement**

The vegetation map for Fort USA was produced from remotely sensed data. The vegetation types were identified using an image classification algorithm applied to the sensed data. For the information derived from this map to be useful in decision making, it should be checked against the physical land features for accuracy, or its' accuracy must be assessed. One way to achieve this is to perform a site-specific error analysis, which compares the remotely sensed data against a "true" (or reference) map of the area. A reference map can be derived from sample data of the area. The LCTA plots were allocated using a stratified random method, which is an appropriate sampling method for accuracy assessment.

## **Acquire Data**

The vegetation types identified on the Fort USA vegetation map will be compared to the plant communities that were determined from LCTA plot data. The Plant Community Classification analysis in the LCTA Program Manager stores data in the LCTA data summary table CommClassPlotSum. This data, along with the vegetation types as defined by the image classification, is extracted from the Fort USA database. Queries for both SQLBase and Access are given. However, if you are using Access LCTA (a program similar to the LCTA Program Manager but written for Microsoft Access) there is no plant community module. The statement given for Access assumes you have imported the CommClassPlotSum table from a SQLBase database.

**SQL Statement (SQLBase syntax)**

```
select plotsurv.plotid, plotsurv.vegtype, commclassplotsum.pccode
from plotsurv, commclassplotsum
where plotsurv.plotid = commclassplotsum.plotid
and @yearno(plotsurv.recdte) = commclassplotsum.analyear
and @yearno(plotsurv.recdte) = 1997
order by plotsurv.plotid;
```

**SQL Statement (Access syntax)**

```
select plotsurv.plotid, plotsurv.vegtype, commclassplotsum.pccode
from plotsurv, commclassplotsum
where plotsurv.plotid = commclassplotsum.plotid
and year(plotsurv.recdte) = commclassplotsum.analyear
and year(plotsurv.recdte) = 1997
order by plotsurv.plotid;
```

The data necessary for the analysis consists of the plot number, vegetation type from remotely sensed data image classification (classified map), and the plant community as determined from a plant community classification (reference map). Any valid plant community classification can be used.

An error matrix is derived from a comparison of the reference map to the classified map using the data described above. Calculated plant communities represent the reference map and form the columns. The classified data from the remotely sensed classified map form the rows. The error matrix is shown below.

Error Matrix

| Classified Data  | Reference Data |               |           |               | Row Marginals |
|------------------|----------------|---------------|-----------|---------------|---------------|
|                  | Dense Woodland | Open Woodland | Grassland | Sparse/Barren |               |
| Dense Woodland   | 30             | 0             | 0         | 0             | 30            |
| Open Woodland    | 3              | 27            | 0         | 0             | 30            |
| Grassland        | 0              | 0             | 30        | 0             | 30            |
| Sparse/Barren    | 0              | 0             | 0         | 20            | 20            |
| Column Marginals | 33             | 27            | 30        | 20            | 110           |

Our vegetation map has only four classifications (dense woodland, open woodland, grassland and sparse/barren). Some of the calculated plant communities were combined to follow the classification of the vegetation map. Dense woodland and closed woodland communities were combined. All grasslands (dense, closed, open, and sparse) were combined as grasslands. We are assuming here that the four classifications of the vegetation map are the ones of importance. For example, we are not interested in discriminating between types of grasslands; we are only interested in the category of grassland as a whole.

The row marginals are the sum of the row values and the column marginals are the sum of the column values. The row marginals represent the number of plots in each classified category. The values in each cell across a row represent the number of plots in the category that fall into the reference data category. For example, the open woodland classified category contains 30 plots, 3 of which were classified as dense woodland using the plant community classification. The remaining 27 plots were classified as open woodland.

## Perform Procedures

We must first determine if there are a sufficient number of reference points (plots), for an overall accuracy assessment of the classification. It has been shown that a minimum sample size of 20 per class is required for 85% classification accuracy, while 30 observations per class are required for 90% accuracy (at the 0.05 confidence level) (Van Genderen and Lock 1977). It should be stated that there are differing ideas on the required sample size per class. Here we will use the values stated above. Notice in the error matrix above, we have sufficient samples for a 90% accuracy assessment for two of four categories (dense woodland and grassland). We have sufficient plots in all four categories for 85% classification accuracy.

Next we should determine the total number of reference points needed to assess the accuracy of the map. The equation below computes the ideal number of points to sample as reference points.

$$N = \frac{Z^2(p)(q)}{E^2}$$

where

N = total number points to be sampled

Z = 2, generalized from the standard normal deviate of 1.96 for the 95% two-sided confidence level

p = expected percent accuracy

q = 100 - p

E = allowable error (standard deviation from the mean)

The total number of points needed for a map with an expected percent accuracy of 85% and an allowable error of 5% is 204.

$$N = (2^2 (85) (15)) / 5^2 \text{ or } 204$$

Fewer sample points can be used if the expected accuracy is assumed to be greater than 85% or if the acceptable standard deviation is larger than 5%. Because we only have 110 plots we do not meet the requirement for an allowable error of 5%. Also, for this example we will assume that our expected accuracy is not greater than 85%. By increasing the acceptable allowable error to 7.5 % we will need approximately 91 plots.

$$N = (2^2 (85) (15)) / 7.5^2 \text{ or } 90.667$$

### ***Percentage of Pixels Correctly Classified***

This is one of the most commonly used measures of agreement and is easy to calculate. Simply divide the number of points correctly classified by the total number of reference points. The equation is shown below.

$$\frac{\sum_{i=1}^r x_{ii}}{\sum_{i=1}^r x_{i+}}$$

The numerator, top value, represents the number of points correctly classified. This value is calculated by summing the diagonal entries from the error matrix. The diagonal values, from upper left to bottom right represent the number of points correctly identified in the classified image as compared to the reference data. The denominator, lower value, is the total number of reference points and is the sum of the row marginals.

From our error matrix above we have:

$$(30 + 27 + 30 + 20) / 110 = .9727 \text{ or } 97.27 \% \text{ of the points were correctly classified.}$$

It is also possible to determine if the percent of correctly classified points exceeds a pre-determined minimum classification accuracy. See Senseman (1995) for further details.

### ***Errors of Omission***

Errors of omission refer to points in the reference map that were classified as something other than their "known" or "accepted" category value. In other words, points of a known category were excluded from that category due to classification error.

Errors of omission for each category are computed by dividing the sum of the incorrectly classified pixels in the nondiagonal entries of that category column by the total number of pixels in that category according to the reference map (the column marginal or total). The values in the

nondiagonal cells represent points that were classified differently in the reference map compared to the classified map.

Calculate the error of omission for dense woodlands from our error matrix. Look down the column of values for dense woodland. Notice the value 3 in the second row under this column. This number represents the number of plots classified as dense woodland, using the plant community classification, that were classified as open woodland on the classified image. The cells in the third and fourth rows contain zero. So, the sum of incorrectly classified points is 3. The value 30 in the first row represents the number of correctly classified points. The error of omission for dense woodlands is computed as:

$$3 / 33 = .0909 \text{ or } 9.09\% \text{ error of omission.}$$

The remaining values were calculated and are shown in the summary table below.

### ***Errors of Commission***

Errors of commission occur when points in the classification map are classified incorrectly and are included in categories in which they do not belong.

Errors of commission are calculated by dividing the sum of incorrectly classified points in the nondiagonal entries of that category row by the total number of points in that category according to the classified map (the row marginal or total).

Calculate the error of commission for open woodland from our error matrix. Read across the row for open woodland. Notice the 3 under the first column. This number represents the number of plots classified as open woodland in the classified map that were classified as dense woodland using the plant community classification. The value in the second column represents the number of plots correctly classified and is not used here. The remaining values in the row are zero, making our sum of incorrectly classified plots 3. The error of commission for open woodland is computed as:

$$3 / 30 = .10 \text{ or } 10\% \text{ error of commission.}$$

The remaining values were calculated and are shown in the summary table below.

### ***Kappa Coefficient of Agreement***

The final measure of agreement discussed is the Kappa Coefficient of Agreement. The Kappa Coefficient measures how well the classification performed compared to the probability of randomly assigning points to their correct categories. The equation for the Kappa Coefficient of Agreement is:

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

where:

- r = the number of rows in the error matrix
- $x_{ii}$  = the number of observation in row i and column i
- $x_{i+}$  = the marginal totals of row i
- $x_{+i}$  = the marginal totals of column i
- N = the total number of observations.

From our error matrix, the Kappa Coefficient is calculated as:

$$\frac{(110*107) - ((30*33) + (30*27) + (30*30) + (20*20))}{110^2 - ((30*33) + (30*27) + (30*30) + (20*20))} = .963$$

It is also possible to calculate a measure of agreement for each class by using the Conditional Kappa Coefficient of Agreement. This is calculated as:

$$K_i = \frac{(N)(p_{ii}) - p_{i+}p_{+i}}{(N)(p_{i+}) - p_{i+}p_{+i}}$$

where:

- $K_i$  = Conditional Kappa Coefficient of Agreement for the ith category
- N = the total number of observations
- $p_{ii}$  = the number of correct observations for the ith category
- $p_{i+}$  = the ith row marginal
- $p_{+i}$  = the ith column marginal.

The equation given in Senseman (1995) failed to add the N terms to the equation.

The Conditional Kappa Coefficient of Agreement for open woodland is:

$$\frac{(110*27) - (27*30)}{(110*30) - (27*30)} = .867$$

## Summary

| Category                 | % Commission | % Omission                | Conditional Kappa |
|--------------------------|--------------|---------------------------|-------------------|
| Dense Woodland           | 0            | 9.09                      | 1                 |
| Open Woodland            | 10           | 0                         | 0.867             |
| Grassland                | 0            | 0                         | 1                 |
| Sparse/Barren            | 0            | 0                         | 1                 |
| <b>Kappa Coefficient</b> |              |                           |                   |
| 0.963                    |              |                           |                   |
| <b>Observed Correct</b>  |              |                           |                   |
| 107                      |              |                           |                   |
| <b>Total Observed</b>    |              | <b>% Observed Correct</b> |                   |
| 110                      |              | 97.3                      |                   |

By examining the measures of agreement in the summary table (above) we can conclude that the classification performed well. 107 of 110 plots (97.3%) were classified correctly. Looking at the values for each of the individual categories we can state that each performed well. The open woodland category was the only one that had plots incorrectly identified in the classification. Three of the open woodland plots were actually classified as dense woodland, using the plant community classification, resulting in a 10% error of commission. This means these plots were included in the classified category of open woodland when they do not belong there. Notice that the dense woodland category has a 9.09% error of omission. This suggests that three plots were classified as something other than their known or accepted category value. In other words, these plots were excluded from dense woodland due to a classification error.

Looking at the Conditional Kappa for each category we conclude all categories, with the exception of open woodland, were accurate. The open woodland was fairly accurate with a value of .8674. The remaining categories were classified correctly.

Keep in mind that this example uses data from a fictitious installation. The LCTA data and imagery of Fort USA are artificial. You should not expect the outcome of an accuracy assessment of your imagery to be quite as good as presented here.

## References

Bishop, Y.M.M., Fienberg, S.E., and P.W. Holland (1975). *Discrete Multivariate Analysis: Theory and Practice*. The MIT Press, Cambridge, MA., 557 pp.

Fitzpatrick-Lins, K. 1981. *Comparison of Sampling Procedures and Data Analysis for a Land-use and Land-cover Map*. Photogrammetric Engineering and Remote Sensing, 47(3): 343-366.

Jensen, John R.. 1986. *Introductory Digital Image Processing*. Prentice-Hall, Englewood Cliffs, NJ., 227 pp.

Senseman, G.M., Bagley, C.F., and S.A. Tweddale. 1995. *Accuracy Assessment of the discrete Classification of Remotely-Sensed Digital Data for Landcover Mapping*. USACERL Technical Report EN-95/04, April 1995.

Van Genderen, J.L., and B.F. Lock. 1977. *Testing Land-Use Map Accuracy*. *Photogrammetric Engineering and Remote Sensing*, 43(9): 1135-1137.