A Coefficient of Agreement as a Measure of Thematic Classification Accuracy

George H. Rosenfield, and Katherine Fitzpatrick-Lins
U.S. Geological Survey, Reston, VA 22092

ABSTRACT: The classification error matrix typically contains tabulated results of accuracy evaluation for a thematic classification, such as a land-use and land-cover map. Diagonal elements of the matrix represent counts correct. The usual designation of classification accuracy has been total percent correct. Nondiagonal elements of the matrix have usually been neglected. A coefficient of agreement is determined for the interpreted map as a whole, and individually for each interpreted category. These coefficients utilize all cell values in the matrix. A conditional coefficient of agreement for individual categories is compared to other methods for expressing category accuracy which have already been published in remote sensing literature.

INTRODUCTION

The classification error matrix typically contains tabulated results of accuracy evaluation for a thematic classification, such as a land-use and land-cover map. Table 1 shows two classification error matrices for two photointerpreters (Congalton and Mead, 1983). In the matrix, for each sample point (or count), interpretation (classification) is given as rows and verification (ground truth) is given as columns.

BACKGROUND

Investigators in remote sensing have been searching for a single value (coefficient) to adequately represent the accuracy of thematic classification. They have also been searching for an accuracy value for each individual category within the map. Because a classification error matrix represents results of accuracy evaluation, the usual procedure has been to use total percent correct—the ratio of the sum of diagonal values to total number of cell counts in the matrix—as map accuracy. Proportions of diagonal values to row sums are considered as category accuracy relative to errors by commission, and proportions of diagonal values to column sums as category accuracy relative to errors of omission.

Development of percent correct, both for the entire map and for individual categories, has been lost to antiquity. Several researchers: (Turk, 1979; Hellden, 1980; Short, 1982) have attempted to find an accuracy value for individual categories which considers errors by commission and errors of omission. Another way of considering overall accuracy and individual category accuracy is to address the problem as a measure of agreement between classification and verification. One measure of agreement that recently received attention in remote sensing applications (Congalton, 1980; Congalton et al., 1981; and Congalton and Mead, 1983) is Cohen’s Kappa coefficient of agreement (Cohen, 1960). This statistic has been given attention by Bishop et al. (1975, p. 395-400). Landis and Koch (1977) define this same situation as “the measure of observer agreement for categorical data.”

PURPOSE AND SCOPE

This paper introduces (1) background development of the Kappa coefficient of agreement as a measure of total map accuracy and (2) the coefficient of conditional agreement (conditional Kappa) as a measure of individual category accuracy.

Development of the Kappa coefficient, its variance, and tests for significant differences are briefly reviewed. Kappa coefficient values and variances are computed for two photointerpreter error matrices (Congalton and Mead, 1983), compared with percent correct, and tested for significant difference. Values of conditional Kappa, its variance, and percent correct are compared.

Accuracy indices (coefficients) for individual categories published in remote sensing literature are computed and compared with each other and with conditional Kappa. Values for percent correct are also computed for each category. Equations for these indices are modified into terms similar to conditional Kappa so they can be compared.
TABLE 1. CLASSIFICATION ERROR MATRICES FOR PHOTOINTERPRETERS #1 AND #2

<table>
<thead>
<tr>
<th>Reference Data</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Total</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35</td>
<td>14</td>
<td>11</td>
<td>1</td>
<td>61</td>
<td>pine</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>11</td>
<td>5</td>
<td></td>
<td>18</td>
<td>cedar</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>9</td>
<td>38</td>
<td>4</td>
<td>63</td>
<td>oak</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>5</td>
<td>12</td>
<td>2</td>
<td>21</td>
<td>cottonwood</td>
</tr>
<tr>
<td>Total</td>
<td>53</td>
<td>39</td>
<td>64</td>
<td>7</td>
<td>163</td>
<td></td>
</tr>
</tbody>
</table>

TABLE 2. CONDITIONAL AGREEMENT FOR METHOD OF TURK

<table>
<thead>
<tr>
<th>Ground Truth (matrix transposed from Turk (1979))</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>148</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>151</td>
<td>other</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>50</td>
<td>6</td>
<td></td>
<td>57</td>
<td>slight and mild</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>15</td>
<td>39</td>
<td>6</td>
<td>68</td>
<td>moderate</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>3</td>
<td>7</td>
<td>25</td>
<td>38</td>
<td>severe</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>8</td>
<td>19</td>
<td>very severe</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>159</td>
<td>68</td>
<td>54</td>
<td>33</td>
<td>322</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Congalton and Mead, 1983, p. 72.

TABLE 3. COMPARISON OF DIFFERENT METHODS FOR EXPRESSING CLASSIFICATION ACCURACY OF INDIVIDUAL CATEGORIES

<table>
<thead>
<tr>
<th>Category</th>
<th>Short</th>
<th>Turk</th>
<th>Helden</th>
<th>( \hat{K}_c \times 100 )</th>
<th>% correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>other</td>
<td>91.36%</td>
<td>92.80%</td>
<td>95.48%</td>
<td>96.08%</td>
<td>98.01%</td>
</tr>
<tr>
<td>slight &amp; mild</td>
<td>66.67%</td>
<td>70.54%</td>
<td>80.00%</td>
<td>84.43%</td>
<td>87.72%</td>
</tr>
<tr>
<td>moderate</td>
<td>46.99%</td>
<td>15.05%</td>
<td>63.93%</td>
<td>48.76%</td>
<td>57.35%</td>
</tr>
<tr>
<td>severe</td>
<td>54.35%</td>
<td>71.01%</td>
<td>70.42%</td>
<td>61.88%</td>
<td>65.79%</td>
</tr>
<tr>
<td>very severe</td>
<td>60.00%</td>
<td>74.41%</td>
<td>75.00%</td>
<td>74.36%</td>
<td>75.00%</td>
</tr>
</tbody>
</table>

A coefficient of agreement as a measure of accuracy

Cohen (1960) developed a coefficient of agreement (called Kappa) for nominal scales which measures the relationship of beyond chance agreement to expected disagreement. This measure of agreement uses all cells in the matrix, not just diagonal elements. The estimate of Kappa \( \hat{K} \) is the proportion of agreement after chance agreement is removed from consideration: that is,

\[
\hat{K} = \frac{(p_o - p_e)}{(1 - p_e)}
\]

in which

- \( p_o = \text{proportion of units which agree} \)
- \( p_e = \text{proportion of units for expected chance agreement} \)
- \( p_o = \frac{\sum \sum p_{ij}}{\sum \sum p_{ij}} \)

where \( N = \text{total number of counts in matrix} \) and \( X_{ij} = \text{number of counts in } i,j \text{th cell} \).

Cohen indicates \( \hat{K} = 0 \) when obtained agreement equals chance agreement. Positive values of Kappa occur from greater than chance agreement; negative values of Kappa are from less than chance agreement. The upper limit of Kappa (+1.00) occurs only when there is perfect agreement. The lower limit of Kappa depends on marginal distributions and is likely to have no practical interest.

Cohen (1960, p. 43) shows how Kappa is similar to product-moment correlation for the dichotomous case. Krippendorf (1970) shows that Kappa belongs to a family of bivariate agreement coefficients, in the form

\[
\text{Agreement} = 1 - \frac{\text{observed disagreement}}{\text{expected disagreement}}
\]

Thus, the coefficient is zero for chance agreement, unity for perfect agreement, and negative for less than chance agreement. Krippendorf shows that several coefficients, including Cohen's Kappa and Pearson's intraclass correlation coefficient (Snedecor and Cochran, 1967, p. 294-296), are related. Furthermore, Fleiss (1975) presents and criticizes a number of proposed indexes of agreement, and stresses the importance of incorporating the value of the index expected by chance alone. He also states (p. 658-659) that Cohen's Kappa coefficient is one of two chance-corrected measures of agreement defensible as intraclass correlation coefficients. The other is Maxwell and Pilliner's coefficient (Maxwell and Pilliner, 1968).

Computation for Cohen's Kappa coefficient treats all disagreements equally. Accordingly, Cohen (1968) developed weighted Kappa to accommodate the
sense of the investigator that some disagreements (non-diagonal cells) are more serious than others. Thus, weighted Kappa describes the proportion of weighted agreement corrected for chance. Equivalence of weighted Kappa with the intraclass correlation coefficient under general conditions is demonstrated by Fleiss and Cohen (1973).

Correct formulation for estimated large sample variances of Kappa and weighted Kappa is given by Fleiss et al. (1969). Earlier versions of variances given by Cohen (1960), Spitzer et al. (1967), Cohen (1968), and Everitt (1968) are incorrect. The approximate large sample variance of Kappa is given by Bishop et al. (1975, p. 396).

Cohen (1960) describes tests of significance between two independent Kappas by

\[ Z = (\hat{K}_1 - \hat{K}_2) / [V(\hat{K}_1) + V(\hat{K}_2)]^{1/2} \]

where \( Z \) is the standard normal deviate. If \( Z \) exceeds 1.96, then the difference is significant at the 95 percent probability level.

A PRACTICAL APPLICATION

Congalton and Mead (1983) utilize Cohen’s Kappa in a practical application to compare results of several photointerpreter classifications. They show classification error matrices for photointerpreters (PI) #1 and #2 (Table 1). Their accuracy analysis was based on “an appropriate sampling scheme” and “an adequate number of samples” (p. 70). Results of their analysis indicate values for Kappa and variance for each photointerpreter, and for the \( Z \) value which measures significance between Kappa values. Values for total percent correct are given for comparison.

\[
\begin{array}{ccc}
\text{Category} & \hat{K} \times 100 & V(\hat{K}) & \% \text{Correct} \\
\hline
\text{pine} & 36.84\% & 0.005821 & 57.37\% \\
\text{cedar} & 48.88 & 0.020743 & 61.11 \\
\text{oak} & 34.66 & 0.006791 & 60.32 \\
\text{cottonwood} & 05.46 & 0.003634 & 9.52 \\
\hline
\end{array}
\]

Again note in each case that values for conditional Kappa are significantly less than values for percent correct.

OTHER ATTEMPTS TO DESCRIBE CATEGORY ACCURACY OF REMOTELY SENSED DATA

METHOD OF TURK (1979)

The ground truth (GT) index developed by Turk (1979) is “the proportion of agreement corrected for chance agreement,” and is given in modified form as

\[ \theta = (X_{ii} - R_i) / (1 - R_i) \]

where \( X_{ii} \) = actual correct classification and \( R_i \) = lucky guesses (chance agreement). \( R_i \) is a derived value.

It is evident that the GT index is also a member of the family of agreement coefficients described by Krippendorf (1970), members of which are obtained by variation in developing observed disagreement and expected frequencies. Turk, as does Cohen (1960), uses the actual correct classification value, \( X_{ii} \). Turk obtains his value for chance agreement, \( R_i \), from a variation of Goodman (1968) based on “quasi-independence” of nondiagonals. Data are fit to the probability model using the method of iterative proportional fitting. Joint occurrence probability is computed as the product of a row parameter and a column parameter, both of which are normalized to unity. Cohen (1960) computes chance agreement...
agreement directly as the product of marginal probabilities.

Table 2 gives the classification error matrix used by Turk (1979, p. 70, Table 2). Turk's matrix has been transposed to have the remote sensing result "on the rows." Turk based this study on the corn leaf blight experiment of Bauer et al. (1971). Note that "very severe" category has only eight sample points, too few to have been considered as a separate category.

METHOD OF HELLDEN (1980)

The mean accuracy (MA) index developed by Hellden (1980, p. 18) "denotes the probability that a randomly chosen point of a specific class on the map has a correspondence of the same class in the same position in the field and that a randomly chosen point in the field of the same class has a correspondence of the same class in the same position on the map," and is given in modified form as

\[
MA = \frac{2X_{ii}}{(X_{ii} + X_{i+})}
\]

where

\[
X_{ii} = \text{number of correctly mapped sample units for specific class,}
\]

\[
X_{i+} = \text{number of sample field units for specific class,}
\]

\[
X_{+i} = \text{number of sampled map units for specific class.}
\]

The MA index is a logical (heuristic) development of Hellden and cannot be derived on either a probability basis or a mathematical basis (correspondence between Rosenfield and Hellden).

METHOD OF SHORT (1982)

The mapping accuracy (MA) index for any class \( i \) developed by Short (1982, p. 259, Table 6-2) is given in modified form as

\[
MA = \frac{X_{ii}}{(X_{ii} + X_{i+})}
\]

This equation of Short is similar to the Jaccard coefficient which Piper (1983) credits to Jaccard (1908). Piper shows that the Jaccard coefficient has the following properties: (1) equals zero if no positive matches, (2) equals one if perfect agreement, and (3) not affected by sample size. In addition, Piper shows that the Jaccard coefficient has the hypergeometric distribution, of which the binomial distribution is a good approximation for large sample sizes.

COMPARISON OF METHODS

To bring the various methods into perspective, Hellden's and Short's methods, along with percent correct and conditional Kappa, were applied to Turk's data matrix. Results of this comparison are given in Table 3.

Note in Table 3 that the order of percentages for each category in general has an upward trend. This observation indicates that percent correct tends to overestimate classification accuracy, while coefficients given by Short, Turk, and Hellden tend to underestimate classification accuracy when compared to conditional Kappa. This observation does not always hold, because coefficients depend upon relative value and location of cell frequency in the classification error matrix.

SUMMARY AND CONCLUSIONS

Kappa coefficient of agreement is shown as a measure of accuracy for thematic classification as a whole, and coefficient of conditional Kappa is presented as an accuracy measure for the individual category. A family of such coefficients exists which correct for chance agreement, but the Kappa coefficient is one of few which also are defensible as intraclass correlation coefficients. These coefficients use information in the classification error matrix resulting from errors by commission and of omission. In the past, the usual designation for accuracy has been percent correct which uses only diagonal elements of the matrix and which appears to give inflated accuracy. Recently, remote sensing literature has given several coefficients as accuracy indices. Only one of these (Turk, 1979) corrects for chance agreement, and none have the statistical basis of being an intraclass correlation coefficient.

It is, therefore, recommended that coefficients of Kappa and conditional Kappa be adopted by the remote sensing community as a measure of accuracy for thematic classification as a whole, and for the individual categories.

REFERENCES


Congalton, R. C., and R. A. Mead, 1983. A quantitative
The Fifth Thematic Conference, to be staged as a part of the continuing series of ERIM Remote Sensing Symposia, will once again address "Remote Sensing for Exploration Geology." As with previous meetings in Ft. Worth (1982), Colorado Springs (1984), and San Francisco (1985), the current 1986-Reno conference will focus upon operational development, emphasizing "Mineral and Energy Exploration: Technology for a Competitive World."

An industry-oriented program, including both conventional plenary sessions and poster presentations, will be formulated to address

- Advanced Sensors and Sensor Systems
- Airborne Remote Sensing Applications
- Applications for Hydrocarbon Exploration
- Applications for Mineral Exploration
- Geobotanical and Environmental Studies
- Global Data Availability/Current and Future
- Multiple Data Sources and Applications
- Photogeology and Image Interpretation

All persons interested in contributing a paper for consideration for poster presentation should submit 30 copies of a comprehensive summary of 200 to 500 words of their proposed presentation and a plan of the poster display, describing the graphics, text, and other materials to be used, no later than 15 March 1986, to

Dr. Jerald J. Cook
Remote Sensing Center
Environmental Research Institute of Michigan
P.O. Box 8618
Ann Arbor, MI 48107-8618
Tele. (313) 994-1200

THEMATIC CLASSIFICATION ACCURACY

• Advanced Sensors and Sensor Systems
• Geobotanical and Environmental Studies
• Airborne Remote Sensing Applications
• Global Data Availability/Current and Future
• Applications for Hydrocarbon Exploration
• Multiple Data Sources and Applications
• Applications for Mineral Exploration
• Photogeology and Image Interpretation

(Rceived 26 May 1984, revised and accepted 26 September 1985)