Statistical Significance and Normalized Confusion Matrices

Perry J. Hardin and J. Matthew Shumway

Abstract

When assessing map accuracy, confusion matrices are frequently statistically compared using kappa. While kappa allows individual matrix categories to be analyzed with respect to either omission or commission error rates, kappa is not used to compare individual matrix categories with respect to both rates concurrently. When this concurrent comparison is desired, the matrices are typically normalized and then scrutinized on a cell-by-cell basis by inspection. While no parametric test of significance exists for such a cell-by-cell examination, sampling distributions for these main diagonal entries can be estimated by repeated subsampling of the original sample data (i.e., bootstrapping), allowing inferences to be made about the population. In this research, the procedure for estimating the sampling distribution of normalized cell values is described. Three methods for determining the standard error of normalized cell value sampling distributions are also outlined. Using these sampling distributions and their attendant standard error, the statistical comparison of cell values from two normalized confusion matrices is illustrated. One illustrated method requires a mild parametric assumption, whereas the other is completely nonparametric. Nevertheless, the two distinct bootstrap methods produce nearly identical results.

Introduction

In remote sensing and geographic modeling, disagreement between nominal maps and reality is frequently tabulated and displayed in a confusion matrix. When multiple classification or modeling methods are used, the resulting confusion matrices are typically compared for significant differences. Because it is one of the few measures which can be tested for significance, Cohens kappa (**K**) (Cohen, 1960) has been the preferred statistic for this confusion matrix comparison.

Recently, however, researchers have been urging caution in the indiscriminate use of **K** without regard to its proper interpretation (Ma and Redmond, 1995) or its correct formulation under stratified sampling schemes (Stehman, 1996). While **K** has been traditionally chosen over other alternatives because it is adjusted for agreement due to random chance alone, Foody (1992) has indicated that **K** is too pessimistic—it underestimates the proportion of agreement by overestimating the random chance component of the concordance.

For detailed confusion matrix analysis, *conditional* kappa (κ) can also be calculated against row or column marginal totals for every matrix class, allowing the accuracy of individual categories to be quantified. κ also allows categories between two confusion matrices to be statistically compared with respect to either actual or predicted class membership (Rosenfield and Fitzpatrick-Lins, 1986). While the κ technique facilitates comparison of individual category error rates with respect to either actual *or* predicted class membership, it cannot be applied with respect to both predicted and actual categories concurrently. In other words, when using κ to discuss class-by-class accuracy, the practitioner must constantly specify whether the context is the predicted or actual class membership rate.

Matrix normalization is another well established confusion matrix analysis procedure (Feinberg, 1970). In contrast to kappa-based methods, matrix normalization provides four principle advantages:

- For any class represented in the normalized matrix, its main diagonal entry provides a single summary measure of the class accuracy with respect to both the predicted and actual marginal totals. Unlike κ, there is no need to refer to the actual or predicted dimension.
- For any class in the normalized matrix, its main diagonal entry takes direct account of both the errors of omission and commission for the class. This incorporation of the off-diagonal cell values is a result of the iterative balancing process which creates the normalized matrix (Congalton *et al.*, 1983).
- Given that the row and marginal totals of normalized confusion matrices sum to a constant, respective cell values in two confusion matrices can be compared directly by inspection.
- When comparing normalized matrices, any two cells in the matrices can be compared. In contrast, κ is limited to the examination of main diagonal cells only.

Statistical significance has been the historical bane of normalized matrix analysis. Normalized cell values have no known parametric sampling distribution; thus, there is no parametric way to determine whether a cell value is significantly different from zero. Furthermore, when contrasting cell values in two normalized matrices, there is no parametric method of determining whether apparent differences are statistically significant—the user is limited to comparison by visual inspection.

Bootstrapping is a Monte Carlo method of estimating a statistic's sampling distribution when a parametric estimator is nonexistent. The bootstrapping process randomly resamples the original sample data many times. For each new sample, the statistic of interest is calculated and recorded. The frequency distribution of the statistic produced from the repetitions is then used as an approximation of the statistic's sampling distribution. After its creation, estimates of standard error (σ_θ) and tests of significance can be derived from the frequency distribution using a variety of methods (Efron and Gong, 1983).

Unless there is some reason for suspecting that the sampling distribution of a statistic is non-normal, there is little reason to determine its standard error by bootstrapping when

Department of Geography, 676 SWKT, Brigham Young University, Provo, UT 84602 (perry_hardin@byu.edu; shumwaym@acd1.byu.edu).

Photogrammetric Engineering & Remote Sensing,
Vol. 63, No. 6, June 1997, pp. 735–740.

0099-1112/97/6306-735$3.00/0

Table 1. Three Methods of Estimating the Standard Error of a Statistic Using a Bootstrapped Sampling Distribution.

Estimate of σ_Θ	Estimation method
Method 1. $\hat{\sigma}_\Theta$	$\hat{\sigma}_\Theta = \sqrt{\frac{\sum_{m=1}^{b} (\Theta_m - \overline{\Theta})^2}{b}}$, where $\overline{\Theta} = \frac{\sum_{m=1}^{b} \Theta_m}{b}$ where Θ_m is the Θ estimate calculated for the mth bootstrap sample. This normal approximation method is suggested by Efron (1979), among many others.
Method 2. $\hat{\sigma}_\Theta$	By counting the sorted values, trim 16 percent from the upper end of the bootstrapped Θ_m distribution. Examine the Θ_m limit of the truncated tail. Designate this upper limit as Θ_{84}. Calculate $\hat{\sigma}_\Theta = (\Theta_{84} - \overline{\Theta})$. This is an elaboration of the percentile confidence interval estimation method described by Mooney and Duval (1993).
Method 3. $\hat{\sigma}_\Theta$	Determine the upper limit of the bias corrected 68 percent confidence interval for the bootstrapped distribution using the method described by Efron (1982). Designate this upper limit as Θ_{84}. Calculate $\hat{\sigma}_\Theta = (\Theta_{84} - \overline{\Theta})$.

a parametric formula is available. However, when the assumption of normality cannot be made, or when statistics such as normalized cell values have no parametric method to determine their standard error, bootstrapping provides an alternative approach (Efron and Tibshirani, 1993).

Problem

Assume that the main diagonal entries for category i in two normalized confusion matrices (${}_1\mathbf{N}$ and ${}_2\mathbf{N}$) are represented by ${}_1a_i$ and ${}_2a_i$ with respective population parameters ${}_1A_i$ and ${}_2A_i$. Can bootstrap methods be applied to the two confusion matrices to estimate the ${}_1A_i$ and ${}_2A_i$ sampling distributions? If so, can ${}_1a_i$ be evaluated against ${}_2a_i$ to determine any statistically significant difference in their magnitude? In this paper, we will demonstrate that both are possible.

After some theoretical background is provided, we present two bootstrapped approaches to comparing ${}_1a_i$ and ${}_2a_i$. The first approach requires a mild parametric assumption, whereas the second requires none. We conclude the paper with a discussion of empty cells in the confusion matrix and how to handle them in the bootstrapping process.

Given the procedures and algorithms described in this paper, matrix normalization becomes a more powerful technique than before. The usual advantages of comparing normalized matrices visually are retained, but now statistical significance for the comparison can be cited as well. The techniques described in this paper also provide a framework for a generalized approach to confusion matrix analysis. Although this paper is limited to describing a method for comparing corresponding main diagonal cells in two matrices, the identical procedure is used to contrast *any* two cells in a pair of confusion matrices.

This paper emphasizes bootstrapping in relation to normalized confusion matrices. While a lengthy discussion of the relative merits of $\mathbf{K}$, κ, and normalization is possible, this paper will not provide that forum. To further focus on methodology, no new data will be presented, and the two confusion matrices published by Congalton and Mead (1983) will be examined.

Theoretical Background

For most descriptive statistical measures such as the arithmetic mean, there are three frequency distributions of interest to the researcher. The first is the population distribution. Unless a complete census has been taken, the population distribution is unknown. The second distribution of interest is the frequency distribution created by sampling the population. From the sample, the researcher traditionally estimates the population parameter by use of a sample statistic. The third distribution is often overlooked, but is implicit in every inferential test—the statistic sampling distribution. Like the population distribution, the sampling distribution is always unknown. For many statistics (e.g., arithmetic mean), theory provides an estimate of its shape. The parameters of the sampling distribution shape (e.g., mean, variance) are determined by theoretically based equations which relate the shape to sample characteristics.

As described by Mooney and Duval (1993), the sampling distribution of any statistic (Θ) "can be thought of as the relative frequency of all possible values of Θ calculated from a sample of [constant and predetermined] size drawn from a given population." As long as a random sample having adequate diversity and size is used, and the function relating the sample characteristics to the sampling distribution is unbiased, the parametric approach to estimating the sampling distribution can be effective. However, it is important to remember that a distribution estimated using the parametric approach is only an approximation to the true, unknown distribution (DiCiccio and Romano, 1989).

Bootstrapping as described by Efron and Gong (1983) is an entirely different approach to estimating the sampling distribution for Θ. Only possible since the advent of modern computers, bootstrapping is simple in its general form:

(1) Randomly extract a sample of size n from the population using an appropriate sampling strategy.
(2) Extract b random samples from the original data sample. Each sample should also be of size n. It is also critical that this sampling be done with replacement. The value of b should be very large ($b > 200$).
(3) Determine Θ for each of the b samples. Sort the set of Θ, creating a distribution. This distribution is an estimate of the sampling distribution for Θ.

Once the sampling distribution for Θ is created, it can be used to make inferences about Θ's corresponding population parameter. The inferential approach chosen depends on researcher preference and analysis goals. In some instances, the distribution itself can be visually inspected to determine the significance of Θ. Alternatively, the standard deviation of the bootstrapped distribution can replace Θ's parametric standard error in traditional inferential tests (e.g., Z-test, t-test). Three methods to determine this standard error are outlined in Table 1.

One of the earliest conceptual discussions of bootstrapping can be found in Efron (1981). Bootstrapping has been used in a variety of physical and social science disciplines. These are reviewed by Mooney and Duval (1993) in a short, easily read volume with a valuable bibliography covering the subject through 1992. Bootstrapping is not entirely new to geography. The use of bootstrapping to validate climatic and other geophysical models is described by Willmott *et al.* (1985).

Bootstrapping Normalized Matrices

As discussed in Congalton *et al.* (1983), a_i can be used to contrast class accuracy in two confusion matrices. The data displayed in Table 2 appeared in Congalton and Mead (1983), and was chosen as fodder for this demonstration. Table 3 is the normalized version of Table 2. Table 4 is the normalized version of a second confusion matrix presented in the same paper. Consider a test for significant difference which could be conducted on the oak category (a_{oak}). As the matrices show, ${}_1a_{oak} = 0.376$ and ${}_2a_{oak} = 0.427$. Is the apparent difference statistically significant, or just a result of random sampling? Representing the population parameters

Table 2. The Original Confusion Matrix Adapted from Congalton and Mead (1983). In the Matrix Normalization Experiments, the Focus Is on the Oak Category.

	Actual Class				
Predicted Class	Pine	Cedar	Oak	Cottonwood	Total
Pine	35	4	12	2	53
Cedar	14	11	9	5	39
Oak	11	3	38	12	64
Cottonwood	1	0	4	2	7
Total	61	18	63	21	

Table 3. Normalized Version of Table 2. Rows and Columns May Not Sum to 1.0 Due to Rounding.

	Actual Class				
Predicted Class	Pine	Cedar	Oak	Cottonwood	Total
Pine	0.512	0.232	0.176	0.080	1.000
Cedar	0.209	0.485	0.121	0.184	0.999
Oak	0.142	0.150	0.376	0.331	0.999
Cottonwood	0.137	0.133	0.326	0.404	1.000
Total	1.000	1.000	0.999	0.999	

corresponding to ${}_1a_{oak}$ and ${}_2a_{oak}$ for the two matrices by ${}_1A_{oak}$ and ${}_2A_{oak}$, the following test can be conducted:

- Level of measurement: Nominal frequencies, normalized
- Model: Random sampling, population characteristics unknown
- Null hypothesis: ${}_1A_{oak} = {}_2A_{oak}$
- Research hypothesis: ${}_2A_{oak} > {}_1A_{oak}$
- Test statistic: Z-test
- Rejection region: $Z_{rej} = -1.64$ (one-tailed, $\alpha = 0.05$)

The closed form formula for a two-sample Z-test is common knowledge. In the context of this problem, substitution produces

$$Z = \frac{{}_2a_{oak} - {}_1a_{oak}}{\sqrt{{}_1\hat{\sigma}^2_{oak} + {}_2\hat{\sigma}^2_{oak}}}, \tag{1}$$

where ${}_1\hat{\sigma}_{oak}$ is the standard error associated with ${}_1a_{oak}$ and ${}_2\hat{\sigma}_{oak}$ is the standard error associated with ${}_2a_{oak}$. The proper use of this formula depends on whether the sampling distributions of ${}_2a_{oak}$ and ${}_1a_{oak}$ are normal, and whether the standard errors are correctly estimated. Neither of these preconditions can be satisfied using any known theoretical assumption or closed formula. However, the following bootstrap process can be used to estimate ${}_1\hat{\sigma}_{oak}$ and ${}_2\hat{\sigma}_{oak}$, as well as to verify the requisite assumptions. These steps must be performed on both matrices independently:

(1) Convert the original confusion matrix into a list of records where the number of records is equal to n. The number of list records for each cell in the matrix is equal to its original cell count c_{ij}. Each record contains a row and column indicator showing the matrix cell which owns it.
(2) Extract a random sample (with replacement) of size n from the list. Using the row and column indicators, constitute a new matrix.
(3) Adjust the matrix for any empty cells. This adjustment is discussed in the next section.
(4) Normalize the matrix using the method established by Feinberg (1970).
(5) Extract a_{oak} from the normalized matrix. Record the value.
(6) Repeat steps 2 through 5 1000 times (b = 1000).
(7) Sort the recorded a_{oak} values and display them as a histogram. This is the estimate of the sampling distribution for the statistic.
(8) Use a Kolmogorov-Smirnov (K-S) test to verify that the sampling distribution is normal.
(9) Estimate $\hat{\sigma}_{oak}$ using the three methods listed in Table 1. Should the K-S test indicate that the sampling distribution is not normal, Method 1 should probably be avoided.

Once the $\hat{\sigma}_a$ estimates for the oak category have been estimated, the Z-test can be applied. However, because there are three estimates for $\hat{\sigma}_{oak}$, the choice of one in preference to the other alternatives requires consideration. Mooney and Duval (1993) describe the relative merits of each approach. Unless there is a reason for preferring one estimate of $\hat{\sigma}_{oak}$ over the other two, the median $\hat{\sigma}_a$ of the three might be used. We usually conduct the Z-test using all nine possible pairs of $\hat{\sigma}_a$ from the two matrices. If the nine Z-tests agree, then the null hypothesis can either be rejected or accepted[1] without worry.

Figure 1 shows the a_{oak} sampling distributions produced from 1000 bootstrap iterations for the two matrices. The K-S test tends to support the normality hypothesis[2] for both curves (In both cases, $d = 0.032$, $p = 0.27$, where d and p denote, respectively, the K-S statistic and its significance). The results of the nine possible Z-tests are summarized in Table 5. Each of the Z values and their associated probabilities (p) are shown. The table also shows the standard error produced for each method outlined in Table 1. Although the different combinations of variance estimates produce different Z-statistics, the null hypothesis is never rejected, i.e., the Z-statistic never approaches the boundary of the rejection region. All the Z-statistics are within 6.6 percent of one another. From these results, it follows that the oak category of the second matrix is not classified significantly better than the oak category of the first matrix. The apparent difference between ${}_1a_{oak}$ and ${}_2a_{oak}$ can be attributed to random sampling. Unlike κ, there is no need to conditionally reference either the a_{oak} statistics or the Z-test against predicted or actual column totals. The normalizing process accounts for both.

The use of Equation 1 in the test above required the minor parametric assumption that the two sampling distributions be normally distributed. The K-S test suggested the assumption was met. However, on occasions when the deviation from normality is severe or the three different estimates of standard error disagree, a completely nonparametric alternative method should be adopted. Again, using the oak cate-

[1]In statistical parlance, the phrase "not rejected" is technically more precise. In this paper, the term "accepted" is used as a synonym to avoid the awkward double negative.

[2]When verifying the assumption of normality prior to conducting a parametric test, typical significance levels such as 0.1 and 0.05 used to reject the null hypothesis (of normality) may be too generous. In this research, we chose to reject the null hypothesis if the p value associated with the K-S test was 0.25 or smaller (closer to zero). In our daily practice, if the null is rejected, we employ one of the nonparametric methods for estimating standard error.

Table 4. Normalized Version of a Second Confusion Matrix. The Original Confusion Matrix Appeared in Congalton and Mead (1983). Rows and Columns May Not Sum to 1.0 Due to Rounding.

	Actual Class				
Predicted Class	Pine	Cedar	Oak	Cottonwood	Total
Pine	0.397	0.295	0.127	0.182	1.001
Cedar	0.226	0.370	0.150	0.254	1.000
Oak	0.061	0.171	0.427	0.341	1.000
Cottonwood	0.317	0.164	0.297	0.223	1.001
Total	1.001	1.000	1.001	1.000	

gory from the two original confusion matrices as an example, the test can be generalized using the following approach:

- Level of measurement: Nominal frequencies, normalized
- Model: Random sampling, population characteristics unknown
- Null hypothesis: ${}_1A_{oak} = {}_2A_{oak}$
- Research hypothesis: ${}_2A_{oak} > {}_1A_{oak}$
- Test statistic: Manual inspection of the sampling distribution of $d_{2\text{-}1}$ created in 1000 bootstrap iterations (b = 1000). The statistic

$$d_{2\text{-}1} = {}_2a_{oak} - {}_1a_{oak}. \qquad (2)$$

- Rejection region: one-tailed, α = 0.05

The bootstrap procedure is slightly more complicated than the previous example:

(1) Convert the first original confusion matrix into a list of records where the number of records is equal to n, i.e., the total of all the cell values of the confusion matrix. The number of list records for each cell in the matrix is equal to its original cell count c_{ij}. As before, each record contains a row and column indicator showing the matrix cell it corresponds to.
(2) Extract a random sample of size n from the first list with replacement. Using the row and column indicators, constitute a new matrix. Adjust the matrix for any empty cells (see the next section).
(3) Normalize the matrix created from the first list.
(4) Extract a_{oak} from the normalized matrix. Record the value.
(5) Repeat steps 2 through 4 1000 times (b = 1000).
(6) Repeat step 1 for the second matrix.
(7) Repeat steps 2 through 4 b times for the second matrix, extracting and recording a_{oak} for each of the b trials.
(8) For each of the b trials performed, calculate $d_{2\text{-}1}$ and record it.
(9) Sort the recorded $d_{2\text{-}1}$ values and display them as a histogram. This is the estimate of the sampling distribution for $d_{2\text{-}1}$.
(10) Count the number of trials where $d_{2\text{-}1} < 0$. Designate this frequency as ξ.
(11) Reject the null hypothesis if ξ is less than $b\alpha$.

Figure 2 shows the distribution of $d_{2\text{-}1}$ produced from following the eleven steps described above. Inspecting the tail of the histogram divulges that 315 of the 1000 $d_{2\text{-}1}$ values are less than or equal to zero. As before, the null hypothesis

Table 5. A Comparison of the Z-Test Results Using Different Standard Error Estimates. Regardless of the Method Used to Estimate Standard Error, the p Value Remains Relatively Unaffected.

	Matrix 1 (Table 3.)		
Matrix 2 (Table 4)	Method 1. ${}_1a_{oak}$ = 0.376 ${}_1\hat{\sigma}_{oak}$ = 0.00511	Method 2. ${}_1a_{oak}$ = 0.376 ${}_1\hat{\sigma}_{oak}$ = 0.00473	Method 3. ${}_1a_{oak}$ = 0.376 ${}_1\hat{\sigma}_{oak}$ = 0.00444
Method 1. ${}_2a_{oak}$ = 0.427 ${}_2\hat{\sigma}_{oak}$ = 0.00491	Z = 0.509 p = 0.305	Z = 0.519 p = 0.302	Z = 0.527 p = 0.299
Method 2. ${}_2a_{oak}$ = 0.427 ${}_2\hat{\sigma}_{oak}$ = 0.00475	Z = 0.514 p = 0.304	Z = 0.524 p = 0.300	Z = 0.532 p = 0.297
Method 3. ${}_2a_{oak}$ = 0.427 ${}_2\hat{\sigma}_{oak}$ = 0.00649	Z = 0.474 p = 0.318	Z = 0.482 p = 0.315	Z = 0.488 p = 0.313

is accepted (ξ = 315, $b\alpha$ = 50, p = 0.315). The probability from this experiment also agrees very favorably with the probability values determined with the Z-test (Table 5) ($0.297 \le p \le 0.318$).

Empty Confusion Matrix Cells

Contrary to popular belief, matrix normalization is not a completely objective procedure. Empty cells in the confusion matrix are usually assigned some arbitrary value k by the researcher in order for matrix normalization to converge. The cell values (a_i) produced in matrix normalization are partially dependent on the k chosen (0.5 and 1.0 are popular). Another selection heuristic is $k = 1/r$, where r represents the number of categories displayed in the matrix. In the oak example described earlier, this value would be 0.25.

As described by Fienberg and Holland (1970) and reviewed recently by Zhuang *et al.* (1995), zero cells in a contingency table can be either fixed or random. Fixed zeros indicate natural impossibilities whereas random zeros result from small or inadequate sample sizes. The empty cells in confusion matrices qualify as random zeros. Fienberg and Holland (1970) recommend three procedures suitable for adjusting a confusion matrix with empty cells. The interested

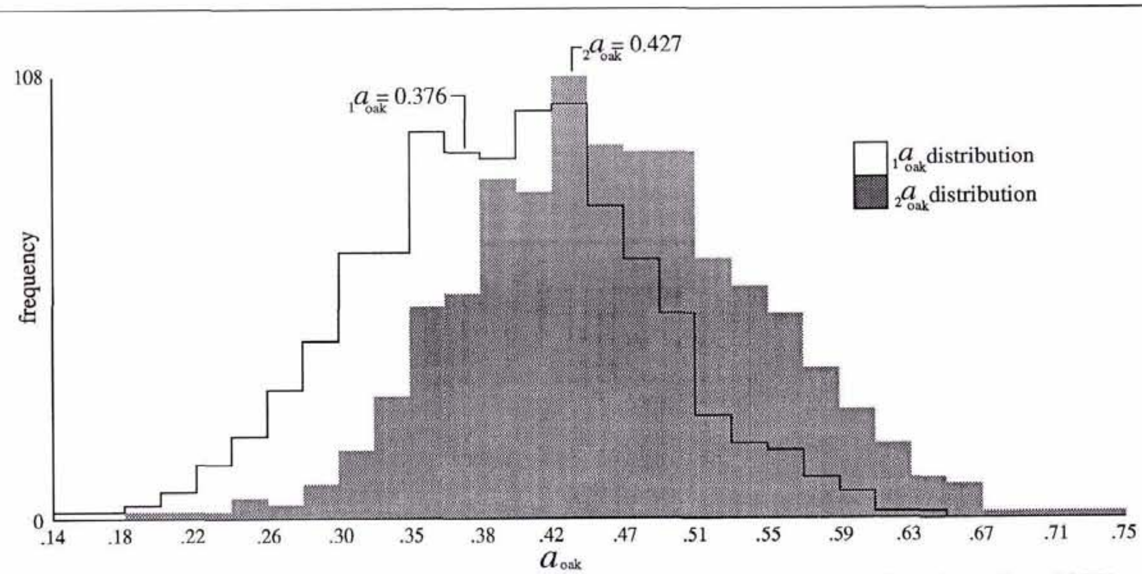

Figure 1. The estimated A_{oak} sampling distributions produced from 1000 bootstrap iterations (b = 1000) for the two normalized matrices. A K-S test suggests that both curves are normal.

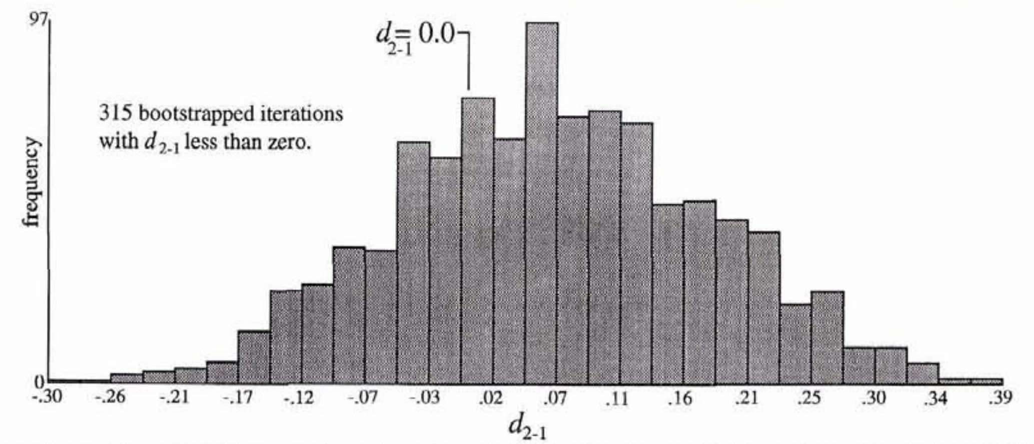

Figure 2. The distribution of $d_{2\text{-}1}$. Inspection of the histogram tail divulges that 315 of the 1000 $d_{2\text{-}1}$ values are less than or equal to zero. Given the statistical criterion, the null hypothesis is not rejected (ξ = 315, $b\alpha$ = 50, p = 0.315).

reader is referred to that article for the theoretical justification and comparison of the three methods. The method used by Zhuang *et al.* (1995) is described by Fienberg and Holland (1970) as the "shrink the matrix to its independent projection" approach. This was the method used to adjust empty confusion matrix cells in this research. It can be described in the following algorithmic form:

(1) Designate the original confusion matrix as **M** with cells m_{ij} where i and j are used to index rows and columns, respectively.

(2) Create a new matrix **E** with cells e_{ij}. The value for any cell e_{ij} can be determined by

$$e_{ij} = m_{+j} \times m_{i+} \div n, \tag{3}$$

where m_{+j} is the marginal total for column j, and m_{i+} is the marginal total for row i. Matrix **E** contains the expected cell counts for the original confusion matrix under the assumption of independence.

(3) Determine the number of pseudo-counts v to be distributed among the cells in the expected confusion matrix **E** by using the formula

$$v = \frac{n^2 - \sum_{i=1}^{r}\sum_{j=1}^{r} m_{ij}^2}{\sum_{i=1}^{r}\sum_{j=1}^{r} (e_{ij} - m_{ij})^2}, \tag{4}$$

where r is the rank of the confusion matrix.

(4) Designate a pseudo-count matrix **P** with cell values p_{ij}. Allocate the v pseudo-counts to matrix **P** using the rule

$$p_{ij} = e_{ij} \times v \div n \tag{5}$$

(5) Add **P** to **M** on a cell-wise basis. After the addition is complete, multiply every cell in **M** by the ratio $n \div (n + v)$ to preserve the original table total of n.

Conclusions

Historically, **K** and κ have been the statistics of choice when comparing two confusion matrices for significant differences. **K** is appropriate when comparing two complete tables, and κ is useful for comparing categories with respect to row or column marginal totals. In contrast, matrix normalization has been recommended when matrix cell values in two tables need to be compared directly. However, unlike κ, no theoretical process for comparing normalized cell values (a_i) for significant differences has existed. Bootstrapping provides a method of assessing the statistical significance of these normalized cells.

In this research, we demonstrated three methods for calculating the standard error of a_i. We also presented two methods of comparing normalized matrix cells directly for statistical significance. One method required the assumption that the sampling distribution be normal. The second method required no such assumption.

Although this demonstration has been limited to comparing values in the main diagonal of confusion matrices, any other pairwise comparison between cells in two matrices can be conducted using the same bootstrapping method. In this paper, several alternative questions regarding the two test matrices could have been addressed. For example, is the error commission rate of the oak-cedar category of the first matrix equal to the error omission rate for the cedar-oak category of the second matrix?

This discussion has also been limited to bootstrapping a_i. Other confusion matrix heuristics seldom reported in the literature can also be assessed for significance using bootstrapping methods. These include the method of Turk(1979), Hellden(1980), and Short(1982).

References

Cohen, J., 1960. A Coefficient of Agreement for Nominal Scales, *Educational and Psychological Measurement*, 20:37–46.

Congalton, R.G., and R.A. Mead, 1983. A Quantitative Method to Test for Consistency and Correctness in Photointerpretation, *Photogrammetric Engineering & Remote Sensing*, 49:69–74.

Congalton, R.G., R.G. Oderwald, and R.A. Mead, 1983. Assessing Landsat Classification Accuracy Using Discrete Multivariate Analysis Statistical Techniques, *Photogrammetric Engineering & Remote Sensing*, 49:671–1687.

DiCiccio, T.J., and J.P. Romano, 1989. The Automatic Percentile Method: Accurate Confidence Limits in Parametric Models, *Canadian Journal of Statistics*, 17:155–169.

Efron, B., 1979. Bootstrap Methods: Another Look at the Jackknife, *Annals of Statistics*, 7:1–26.

———, 1981. Nonparametric Standard Errors and Confidence Intervals, *Canadian Journal of Statistics*, 9:139–172.

———, 1982. *The Jackknife, the Bootstrap, and other Resampling Plans*, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania. As quoted in Mooney and Duval (1993, pp. 37–40).

Efron, B., and G. Gong, 1983. A Leisurely Look at the Bootstrap, the Jackknife and Cross-Validation, *American Statistician*, 37:36–48.

Efron, B., and R. Tibshirani, 1993. *An Introduction to the Bootstrap*, Chapman and Hall, New York.

Feinberg, S.E., 1970. An Iterative Procedure for Estimation in Contingency Tables, *Annals of Mathematical Statistics*, 41:907–917.

Feinberg, S.E., and P.W. Holland, 1970. Methods for Eliminating Zero Counts in Contingency Tables, *Random Counts in Scientific Work* (G.P. Patil, editor), Pennsylvania State University, University Park, Pennsylvania, 1:233–260.

Foody, G.M., 1992. On the Compensation for Change Agreement in Image Classification Accuracy Assessment, *Photogrammetric Engineering & Remote Sensing*, 58:1459–1460.

Hellden, U., 1980. *A Test of Landsat-2 Imagery and Digital Data for Thematic Mapping Illustrated by an Environmental Study in Northern Kenya*, Lund University Natural Geography Institute Report No. 47, Sweden.

Ma, Z., and R.L. Redmond, 1995. Tau Coefficients for Accuracy Assessment of Classification of Remotely Sensed Data, *Photogrammetric Engineering & Remote Sensing*, 61:435–439.

Mooney, C.Z., and R.D. Duval, 1993. *Bootstrapping: A Nonparametric Approach to Statistical Inference*, Sage Publications, Newbury Park, California.

Rosenfield, G.H., and K. Fitzpatrick-Lins, 1986. A Coefficient of Agreement as a Measure of Thematic Classification Accuracy, *Photogrammetric Engineering & Remote Sensing*, 52:223–227.

Short, N.M., 1982. *The Landsat Tutorial Workbook - Basics of Satellite Remote Sensing*, NASA Reference Publication 1078, Goddard Space Flight Center, Greenbelt, Maryland.

Stehman, S.V., 1996. Estimating the Kappa Coefficient and its Variance Under Stratified Random Sampling, *Photogrammetric Engineering & Remote Sensing*, 62:401–407.

Turk, G., 1979. GT Index: A Measure of the Success of Prediction, *Remote Sensing of Environment*, 8:65–75.

Willmott, C.J., S.G. Ackleson, R.E. Davis, J.J. Feddema, K.M. Klink, D.R. Legates, J. O'Donnell, and C.M. Rowe, 1985. Statistics for the Evaluation and Comparison of Models, *Journal of Geophysical Research*, 90(C5):8995–9005.

Zhuang, X., B.A. Engel, X. Xiong, and C.J. Johannsen, 1995. Analysis of Classification Results of Remotely Sensed Data and Evaluation of Classification Algorithms, *Photogrammetric Engineering & Remote Sensing*, 61:427–433.

(Received 20 May 1996; accepted 16 August 1996; revised 9 January 1997)