Thematic Map Comparison: Evaluating the Statistical Significance of Differences in Classification Accuracy

Giles M. Foody

Abstract

The accuracy of thematic maps derived by image classification analyses is often compared in remote sensing studies. This comparison is typically achieved by a basic subjective assessment of the observed difference in accuracy but should be undertaken in a statistically rigorous fashion. One approach for the evaluation of the statistical significance of a difference in map accuracy that has been widely used in remote sensing research is based on the comparison of the kappa coefficient of agreement derived for each map. The conventional approach to the comparison of kappa coefficients assumes that the samples used in their calculation are independent, an assumption that is commonly unsatisfied because the same sample of ground data sites is often used for each map. Alternative methods to evaluate the statistical significance of differences in accuracy are available for both related and independent samples. Approaches for map comparison based on the kappa coefficient and proportion of correctly allocated cases, the two most widely used metrics of thematic map accuracy in remote sensing, are discussed. An example illustrates how classifications based on the same sample of ground data sites may be compared rigorously and highlights the importance of distinguishing between one- and two-sided statistical tests in the comparison of classification accuracy statements.

Introduction

Thematic mapping through the process of an image classification is one of the most commonly undertaken analyses in remote sensing. Despite the considerable potential of remote sensing for mapping a range of themes, notably land cover, numerous problems are often encountered that may result in significant error. Because the magnitude of classification error can be large, information on the accuracy of a map is often required. Map users, for instance, require accuracy information to indicate the quality of the map and its suitability for a particular purpose. The map producer also requires information on map accuracy, especially as a means to evaluate and refine the mapping process. An accuracy statement is, therefore, an important accompaniment to any thematic map derived from remote sensing. Although it is now widely accepted that no classification is complete until its accuracy has been rigorously assessed, many problems are often encountered in accuracy assessment (Foody, 2002). This complicates not only the evaluation of a single map but also the comparison of different maps.

Many methods have been used in the assessment of thematic map accuracy. Most commonly in remote sensing, a site specific approach is used in which the predicted and actual class labels for a set of specific sites are compared. Typically, the assessment of map accuracy is based on a confusion or error matrix, a simple cross-tabulation of the predicted and actual class labels for the selected sites (Table 1). A range of metrics to describe the accuracy of a thematic map may be derived from the confusion matrix, with the proportion of correctly classified cases (often expressed as a percentage) and the kappa coefficient of agreement used most frequently in remote sensing (Trodd, 1995). Because an accuracy statement is generally based on the class allocations observed for a sample of sites drawn from the map, it provides, however, only an estimate of the map's accuracy. For this reason, accuracy statements should ideally be accompanied by confidence limits (Thomas and Allcock, 1984; Janssen and van der Wel, 1994) and be compared in a statistically rigorous fashion. However, only rarely are confidence limits provided in published papers. Perhaps more critically, many map comparisons are based on subjective evaluations, commonly involving little more than the direct comparison of the magnitude of the derived estimates of map accuracy (e.g., Foody, 2001; Chen and Stow, 2002; Sohn and Rebello, 2002). Ideally, the comparison of thematic map accuracy should be undertaken statistically, to provide a more objective basis for comment and interpretation.

Map accuracy statements are compared for a variety of reasons, but commonly to gain a relative evaluation of two or more classifications. Often in remote sensing applications, the comparison is undertaken to evaluate the relative suitability of different classification techniques for mapping. Much of this work has been driven by the problems encountered in thematic mapping in the past and aims to increase the accuracy with which thematic maps can be derived from remotely sensed data. Thus, for example, there is a very large literature involving the relative comparison of different classification techniques (e.g., maximum likelihood versus neural networks) or approaches (e.g., per-pixel versus per-parcel classification). In such studies, the key focus is on the difference in the estimated classification accuracies.

Of those studies that have sought to evaluate the statistical significance of differences in classification accuracy, many have based the analysis on the comparison of accuracy expressed in terms of the kappa coefficient of agreement. This situation reflects the widespread use of the kappa coefficient of

School of Geography,University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom (g.m.foody@soton.ac.uk)

Photogrammetric Engineering & Remote Sensing
Vol. 70, No. 5, May 2004, pp. 627–633.

0099-1112/04/7005–0627/$3.00/0

TABLE 1. CONFUSION MATRICES FOR THE FOUR CLASSIFICATIONS PERFORMED IN THE EXAMPLE. THE MAIN DIAGONAL, PRESENTED BOLDFACE IN EACH MATRIX, SHOWS THE CASES CORRECTLY ALLOCATED. NOTE THAT THE SAME SAMPLE OF SITES WAS USED IN THE GENERATION OF EACH MATRIX

1A. DISCRIMINANT ANALYSIS

	Predicted Class						
Class	Wheat	Sugar Beet	Barley	Potato	Carrot	Grass	Σ
Actual class							
Wheat	**77**	11	9	0	1	0	98
Sugar beet	7	**22**	2	20	0	0	51
Barley	2	0	**21**	0	0	0	23
Potato	0	0	0	**12**	0	0	12
Carrot	0	0	0	3	**7**	0	10
Grass	0	0	0	0	4	**2**	6
Σ	86	33	32	35	12	2	200

Discriminant analysis: Kappa coefficient = 0.587; Proportion correct = 0.705.

1B. MLP

	Predicted Class						
Class	Wheat	Sugar Beet	Barley	Potato	Carrot	Grass	Σ
Actual class							
Wheat	**78**	12	8	0	0	0	98
Sugar beet	5	**44**	0	1	1	0	51
Barley	4	2	**17**	0	0	0	23
Potato	0	0	0	**12**	0	0	12
Carrot	0	0	0	0	**9**	1	10
Grass	0	0	0	0	3	**3**	6
Σ	87	58	25	13	13	4	200

MLP: Kappa coefficient = 0.732; Proportion correct = 0.815.

1C. PNN

	Predicted Class						
Class	Wheat	Sugar Beet	Barley	Potato	Carrot	Grass	Σ
Actual class							
Wheat	**81**	13	4	0	0	0	98
Sugar beet	7	**42**	0	1	1	0	51
Barley	5	2	**16**	0	0	0	23
Potato	0	0	0	**12**	0	0	12
Carrot	0	0	0	0	**10**	0	10
Grass	0	0	0	0	4	**2**	6
Σ	93	57	20	13	15	2	200

PNN: Kappa coefficient = 0.728; Proportion correct = 0.815.

1D. PNN WITH PRIOR INFORMATION (EXPRESSED IN TERMS OF THE PROPORTIONAL COVERAGE OF THE CLASSES AT THE TEST SITE)

	Predicted Class						
Class	Wheat	Sugar Beet	Barley	Potato	Carrot	Grass	Σ
Actual class							
Wheat	**88**	7	3	0	0	0	98
Sugar beet	4	**44**	2	0	1	0	51
Barley	5	2	**16**	0	0	0	23
Potato	0	0	0	**12**	0	0	12
Carrot	0	0	0	4	**6**	0	10
Grass	0	0	0	0	4	**2**	6
Σ	97	53	21	16	11	2	200

PNN with prior information: Kappa coefficient = 0.762; Proportion correct = 0.840.

agreement as an accuracy metric in remote sensing and also the explicit discussion of the comparison of kappa coefficients in the literature. Indeed, the ability to compare kappa coefficients has been advocated as a valuable feature since the kappa coefficient was first introduced to the remote sensing community (Congalton and Mead, 1983; Congalton *et al.*, 1983; Rosenfield and Fitzpatrick-Lins, 1986; Janssen and van der Wel, 1994; Smits *et al.*, 1999). There are, however, concerns about the use of kappa coefficient for accuracy assessment and comparison. For instance, the sampling design used to acquire the data contained in a confusion matrix has implications for the estimation of the kappa coefficient (Stehman, 1996). Here,

the focus is on issues connected to the comparison of accuracy statements. Specifically, the aim is to highlight a common problem in accuracy comparison based upon the kappa coefficient and show that alternatives, compatible with popular accuracy assessment methods used in remote sensing, are available. For simplicity, the discussion below ignores a host of complicating factors, assuming, for example, that simple random sampling is used in the evaluation of classifications comprising mutually exclusive and unordered classes in which all misallocations are of equal weight.

Comparison of Kappa Coefficients

The evaluation of the statistical significance of the difference in accuracy between two thematic maps derived from remotely sensed data has often been based on the comparison of the kappa coefficient calculated for each map. The kappa coefficient of agreement for a thematic map is based on the comparison of the predicted and actual class labels for each case in the testing set and may be calculated from

$$\hat{\kappa} = \frac{p_o - p_c}{1 - p_c} \tag{1}$$

where p_o is the proportion of cases in agreement (i.e., correctly allocated) and p_c is the proportion of agreement that is expected by chance. The derived coefficient provides an estimate of the accuracy of the map which together with that derived from another map is the basis of most map comparisons. Specifically, the map comparison seeks to determine if the difference in the derived estimates can be inferred to indicate a difference in the associated population parameters of accuracy. The significance of the difference in accuracy between two maps with *independent* kappa coefficients, $\hat{\kappa}_1$ and $\hat{\kappa}_2$, may be evaluated with the normal curve deviate

$$z = \frac{\hat{\kappa}_1 - \hat{\kappa}_2}{\sqrt{\hat{\sigma}^2_{\kappa_1} + \hat{\sigma}^2_{\kappa_2}}} \tag{2}$$

where $\hat{\sigma}^2_{\kappa_1}$ and $\hat{\sigma}^2_{\kappa_2}$ represent the estimated variances of the derived coefficients. The significance of the difference between the two kappa coefficients is then assessed by comparing the value of z calculated from Equation 2 against tabulated values. For the simple situation of determining if there is a difference between two kappa coefficients (two-sided test), the null hypothesis (H_o), of no significant difference, would be rejected at the widely used 5 percent level of significance if $|z| > 1.96$ (Congalton *et al.*, 1983; Rosenfield and Fitzpatrick-Lins, 1986; Congalton and Green, 1999).

The approach for accuracy comparison based on Equation 2 has been widely used for the comparison of classification accuracy statements in the remote sensing literature. There are, however, some important concerns associated with the use of this approach for the comparison of kappa coefficients. For example, one widely recognized problem in the remote sensing literature is that the equation for calculating the variance provided in Cohen's (1960) paper was incorrect (e.g., Rosenfield and Fitzpatrick-Lins, 1986; Hudson and Ramm, 1987). A further fundamental concern is that the approach for kappa coefficient comparison may be inappropriate in many instances. This is because the assumption of the independence of the samples used, which is highlighted both above and in the original paper by Cohen (1960, p. 44), has often not been satisfied. This is apparent in numerous discussions in the literature spanning the period from the introduction of the kappa coefficient to the remote sensing community through to the present day, in which the samples used in the calculation of the kappa coefficients are matched or related. For example, many studies use the same sample of sites to form the confusion matrices from which kappa coefficients to be compared are derived (e.g., Congalton *et al.*, 1983; Haack *et al.*, 2002; Sohn and Rebello, 2002). In studies evaluating different classification algorithms it is, for instance, common to find that the same set of ground data is used in assessing the accuracy of each classifier to be compared. Because the same sample of data is used in the derivation of each kappa coefficient, the assumption of independence is not satisfied and the approach outlined above should not be used to evaluate the statistical significance of differences in map accuracy indicated by the derived kappa coefficients.

Comparison of Kappa Coefficients for Related Samples

The method of comparing kappa coefficients outlined above should not be used for the comparison of kappa coefficients estimated from the same or related samples (McKenzie *et al.*, 1996). Relatively little attention has been directed at the problem of comparing kappa coefficients derived from related samples (Donner *et al.*, 2000) and it is apparent that there is no parametric test that is suitable for this task (McKenzie *et al.*, 1996). Recently, however, approaches to accommodate for the dependence between kappa coefficients arising through the use of related samples have been proposed. Donner *et al.* (2000), for example, extend the method used for data from independent samples, outlined above, to account for the covariance between kappa statistics due to the use of a related sample. This approach is appropriate for the comparison of kappa coefficients derived from classifications with a dichotomous outcome. Consequently, this approach may be useful for simple two-class classifications, such as in change detection in which the change matrix (Congalton and Green, 1999) upon which the analysis is based shows classes of "change" and "no change." In the method presented by Donner *et al.* (2000), the significance of the difference between two kappa coefficients derived from a related sample is based on

$$z = \frac{\hat{\kappa}_1 - \hat{\kappa}_2}{\left(\hat{\sigma}^2_{\kappa_1} + \hat{\sigma}^2_{\kappa_2} - 2\hat{\sigma}_{\hat{\kappa}_1\hat{\kappa}_2}\right)} \tag{3}$$

where $\hat{\sigma}_{\hat{\kappa}_1\hat{\kappa}_2}$ represents the estimated covariance between $\hat{\kappa}_1$ and $\hat{\kappa}_2$ (Donner *et al.*, 2000).

An alternative approach to comparing kappa coefficients derived from related samples is presented by McKenzie *et al.* (1996) and is based upon the use of resampling techniques. In this method, a large number of samples are derived from the original sample to derive a probability distribution of the statistic. The method discussed by McKenzie *et al.* (1996) uses Monte Carlo permutation tests in the determination of the statistical significance of the difference between two kappa coefficients derived using related samples. For this, the variable in common to the two classifications (e.g., the "actual" class label derived from the ground data in the example given below) are randomly shuffled and the kappa coefficients are recomputed. For each permutation, the difference between the derived kappa coefficients is estimated. Typically, many hundreds or thousands of random permutations are undertaken, with approximately 1000 permutations being adequate for significance testing at the 5 percent level of significance (McKenzie *et al.*, 1996). The number of times that the original difference in the kappa coefficients is equaled or exceeded by the difference in the randomly permuted values derived in the analysis is noted, incremented by 1 and divided by the total number of permutation plus 1 to derive a proportion. Assuming the common situation in which a two-sided test (H_o that there is no significant difference between the two kappa coefficients) at the 5 percent level of significance is being undertaken, the difference between two kappa coefficients derived with related samples would be regarded as being statistically significant if the computed proportion was less than 0.05.

Alternatives Based on the Proportion Correct

There is nothing unique about the kappa coefficient in terms of the ability to statistically compare values. Indeed, given the contentious nature of the kappa coefficient as a metric of classification accuracy and problems with its use in remote sensing (Stehman, 1997; Foody, 2002), it may be preferable to use other metrics for accuracy assessment in remote sensing. Many other metrics may be used to estimate classification accuracy and, contrary to some earlier statements (e.g., Janssen and van der Wel, 1994), these metrics can be compared statistically. Moreover, approaches exist for the comparison of classifications derived with both related and independent samples. Thus, for example, the receiver operating characteristic (ROC) curve may be used to provide a powerful means of accuracy assessment and comparison for analyses based on simple two-class situations (Zweig and Campbell, 1993). Similarly, approaches exist for the comparison of accuracy statements expressed as the proportion correct allocation. Because the proportion of correctly allocated cases (often expressed as a percentage) is perhaps the most commonly used accuracy metric in remote sensing (Trodd, 1995), this approach will be discussed briefly before providing an example of the calculation of some of the metrics for illustrative purposes.

Independent Samples

If independent samples have been used, the significance of the difference between two proportions may be estimated from

$$z = \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (4)$$

where x_1 and x_2 represent the number of correctly allocated cases in two independent samples of size n_1 and n_2, respectively, and $p = (x_1 + x_2)/(n_1 + n_2)$. The estimated proportion of correctly allocated cases for each classification (e.g., x_1/n_1) can be converted to the percentage correct allocation by multiplying by 100 and then used to indicate map accuracy. The statistical significance of the difference in accuracy between two classifications is evaluated through z in the same way as with the comparison of kappa coefficients. This method should only be used when the samples are reasonably large (Freund and Williams, 1959; Fienberg, 1981; Clark and Hosking, 1986; Neter *et al.*, 1993). A slightly different approach to the comparison of proportions has also been suggested by Stehman (1997).

Related Samples

In many remote sensing studies the same set of sites are used in the assessment of the accuracy of the thematic maps to be compared. Consequently, the samples are not independent, and an alternative approach to that outlined above that is suitable for related samples is required. For related samples, the statistical significance of the difference between two proportions may be evaluated using McNemar's test (Bradley, 1968; Agresti, 1996). This is a non-parametric test that is based upon confusion matrices that are 2 by 2 in dimension. The constraint on the size of the matrices is often not a problem because larger matrices can be collapsed to this size because attention is, in effect, focused on the binary distinction between correct and incorrect class allocations. The McNemar test is based upon the standardized normal test statistic

$$z = \frac{f_{12} - f_{21}}{\sqrt{f_{12} + f_{21}}} \quad (5)$$

in which f_{ij} indicates the frequency of sites lying in confusion matrix element i, j (further clarification on this is given in the example below, and in Table 2A in particular). Commonly in the literature, some discussions of this technique, including its use within remote sensing to compare accuracy statements (e.g., Chan *et al.*, 2003), bases the evaluation upon a chi-square (χ^2) distribution; the square of z follows a chi-squared distribution with one degree of freedom (Agresti, 1996). In such circumstances, the test equation may be expressed as

$$\chi^2 = \frac{(f_{12} - f_{21})^2}{f_{12} + f_{21}} \quad (6)$$

with the derived value compared against tabulated chi-squared values to indicate its statistical significance. This approach cannot, however, be used for testing a one-sided hypothesis because the rejection region for the chi-squared test is one-tailed.

TABLE 2. ASSESSMENT OF THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN TWO CLASSIFICATIONS WITH THE MCNEMAR TEST

2A. THE DEFINITION OF MATRIX ELEMENTS USED IN EQUATIONS 5, 6, AND 8

	Classification 2		
Allocation	Correct	Incorrect	Σ
Classification 1			
Correct	f_{11}	f_{12}	
Incorrect	f_{21}	f_{22}	
Σ			

2B. THE COMPARISON OF THE CLASSIFICATIONS DERIVED FROM THE PNN WITH AND WITHOUT PRIOR INFORMATION AS AN EXAMPLE

	PNN		
Allocation	Correct	Incorrect	Σ
PNN with prior information			
Correct	158	10	168
Incorrect	5	27	32
Σ	163	37	200

Continuity Correction

Because a continuous (normal) distribution is being used to represent the discrete distribution of sample frequencies in tests based on z, it is sometimes recommended that a correction for continuity be undertaken (Fleiss, 1981). For this, the test equations expressed by equations 4 and 6 may be modified, with the test for the difference between two proportions derived from independent samples based on

$$z = \frac{\left|\frac{x_1}{n_1} - \frac{x_2}{n_2}\right| - \frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (7)$$

and the McNemar test for related samples on

$$\chi^2 = \frac{(|f_{12} - f_{21}| - 1)^2}{f_{12} + f_{21}}. \quad (8)$$

Continuity correction is particularly important if the sample size used is small (Neter *et al.*, 1993), its impact diminishing at large sample sizes, but the requirement for its use is the subject of debate (e.g., Fienberg, 1981; Fleiss, 1981).

Example

The approaches for accuracy comparison based on independent samples are well-established in the literature, so attention here focuses on the commonly encountered situation in remote sensing in which related samples are used. For this example, a series of supervised classifications was undertaken to

illustrate the comparison of accuracies estimated with a related sample.

Each classification analysis aimed at deriving a thematic map of an agricultural region from three wavebands of data acquired by an airborne thematic mapper sensor. For each class, 100 randomly selected pixels were used to form the training data set. The accuracy of the classifications was assessed from an independent (from the training set) sample of 200 randomly selected pixels. Further details on the data are given in Foody (2001). Thus, in common with many other studies that have sought to evaluate different classifiers, the same training and testing sets were used in each classification, helping to ensure that differences in accuracy could be attributed to the nature of the class allocation processes used.

Three different classifiers were applied to the data, a standard statistical classification using a discriminant analysis and two feedforward neural network classifiers, a standard multi-layer perceptron (MLP), and a probabilistic neural network (PNN). A confusion matrix was generated for each classification and its accuracy, expressed by the kappa coefficient of agreement and proportion of pixels correctly allocated, assessed (Table 1). The statistical significance of the difference between selected accuracy statements was then determined. These evaluations were based on the method outlined by McKenzie *et al.* (1996) with the kappa coefficients and the McNemar test with the proportion of correct allocations.

The resampling method presented by McKenzie *et al.* (1996) was applied to the data illustrated in Table 1 using 9999 permutations. Thus, 9999 random pairings of the actual class labels, with the predicted labels defined by the two classifications to be compared, were undertaken to determine the statistical significance of the difference in accuracy between the classifications indicated by the kappa coefficients. Pairwise comparisons of the classifications were also undertaken using the McNemar test, without correction for continuity, applied to the proportion of correct agreement calculated for the classifications. The results of a selection of the comparisons are summarized in Table 3 and used to answer three questions typical of those posed in remote sensing studies:

Q1. Is there a significant difference in the accuracy of the classifications derived from the discriminant analysis and neural networks? This question may be answered by comparing the accuracy metrics for the classification derived from the discriminant analysis (Table 1A) against those derived from the two neural networks (Tables 1B and 1C). From Table 3, it is apparent that the large differences in accuracy observed between the classifications expressed in terms of both the kappa coefficient of agreement and proportion of correctly allocated cases are statistically significant at the 0.1 percent level of significance.

Q2. Is there a significant difference in the accuracy of the classifications derived from the two neural networks? Although there is evidence that the MLP classification appears to be marginally more accurate than that derived from the PNN (Tables 1B and 1C), the difference is statistically insignificant at the 5 percent level of significance (Table 3).

Q3. Unlike the widely used MLP, it is possible to usefully incorporate prior knowledge into the classification with a PNN, which would be expected to increase classification accuracy. Is there a significant difference in the accuracy of the classifications derived from the PNN with and without prior information of class occurrence? As with the situation in relation to Q2, a straightforward comparison of the accuracy metrics (Tables 1C and 1D) indicates that a difference in accuracy exists but, on further examination, it is found to be statistically insignificant (at the 5 percent level of significance). With the comparison based on the kappa coefficients, for example, the observed difference (0.035) was exceeded by that calculated in 935 of the 9999 random

TABLE 3. SUMMARY OF SOME OF THE CLASSIFICATION COMPARISONS UNDERTAKEN. THE RESAMPLING METHOD WAS USED TO COMPARE THE KAPPA COEFFICIENTS AND THE MCNEMAR TEST TO COMPARE THE PROPORTIONS OF CORRECTLY ALLOCATED PIXELS. ALL TESTS SHOWN WERE TWO-SIDED AND THE 5 PERCENT LEVEL OF SIGNIFICANCE SELECTED, WITH THE LEVEL OF SIGNIFICANCE STATED FOR SIGNIFICANT DIFFERENCES

		Comparison of Kappa Coefficients				Comparison of Proportions				
Classification 1	Classification 2	$\hat{\kappa}_1$	$\hat{\kappa}_2$	$\hat{\kappa}_1 - \hat{\kappa}_2$	Significant?	$\frac{x_1}{n_1}$	$\frac{x_2}{n_2}$	$\frac{x_1}{n_1} - \frac{x_2}{n_2}$	$\lvert z \rvert$	Significant?
Discriminant analysis	MLP	0.587	0.732	−0.144	Yes, 0.1%	0.705	0.815	−0.110	3.88	Yes, 0.1%
Discriminant analysis	PNN	0.587	0.728	−0.141	Yes, 0.1%	0.705	0.815	−0.110	3.77	Yes, 0.1%
MLP	PNN	0.732	0.728	0.004	No, 5%	0.815	0.815	0.000	0	No, 5%
PNN & prior information	PNN	0.763	0.728	0.035	No, 5%	0.840	0.815	0.025	1.29	No, 5%

permutations undertaken, yielding a proportion of 0.0936 (935 + 1/9999 + 1). This result indicates that the observed difference is insignificant at the 5 percent level of significance. However, because the incorporation of prior information into the analysis would be expected to *increase* classification accuracy, rather than simply causing a change in accuracy, it may be more appropriate to restate Q3 to indicate the direction of change. This would then allow a one- rather than two-sided test to be undertaken. Thus, re-expressing the question in terms of the incorporation of prior information increasing classification accuracy, the alternative hypothesis to H_o becomes, essentially, that the classification derived from the PNN with prior information is more accurate than that derived without prior information. In this situation, the comparison of the kappa coefficients indicates that the observed difference in accuracy is significant at the 5 percent level of significance; the weaker analysis based on the proportion correct did not identify a significant difference at the 5 percent level of significance. Finally, it is important to note that the interpretation derived may differ from that arising through the (inappropriate) use of the techniques that assume independent samples. For example, if the approach based on Equation 2 had been (inappropriately) used, the analyst would have been content to uphold the H_o of no difference for both one- and two-sided tests as the calculated value of z, 0.63, was substantially below the critical value.

In relation to each of the three questions above, a simple subjective assessment could be that a significant difference in accuracy exists (as a non-zero difference was observed in each case), and so by inference that the classifiers vary in their utility for thematic mapping. A more objective assessment of the differences is, however, provided by the statistical testing described which may result in different interpretations of the results and thereby of the relative utility of the classifiers. For example, while the large difference in accuracy between the discriminant analysis and neural networks is statistically significant, that observed between the two neural networks is not. Finally, the example demonstrated the importance of using, if appropriate, a one- rather than two-sided test. Thus, if the question posed has a sensible directional component, this should be considered in the testing because this can, as in the example, result in a different interpretation being drawn.

Summary and Conclusions

There is often a desire to compare the accuracy of different classifications. One approach that has been used widely in remote sensing is to undertake a pairwise comparison of the accuracy statements, based on the kappa coefficient of agreement, derived for each classification. This analysis is based on the widely promoted approach represented by Equation 2. While this may sometimes be an appropriate approach for accuracy comparison, it does assume that the kappa coefficients under comparison are independent. Frequently, the assumption of independent samples is unsatisfied. Often the same sample of data has been used in the derivation of the confusion matrices from which the kappa coefficients were derived. Indeed, the use of related samples is common in remote sensing, particularly in research that aims to evaluate the relative accuracy of different image classifiers or mapping approaches. Therefore, many researchers, including the present author, have inappropriately sought to evaluate the statistical significance of differences in accuracy statements derived from related samples with the use of a technique that assumes independent samples. Although the error arising from this situation may sometimes be small and perhaps does not change the fundamental interpretations drawn, such practices should be discouraged. Moreover, the effect of mis-using a technique can be large, and, as evident in the example presented, can result in mis-interpretation of the significance of a difference in accuracy. This is particularly important because appropriate alternative means of accuracy comparison exist for a range of accuracy metrics derived from related samples. The example presented above, for instance, showed that, using a resampling method, it is possible to rigorously compare accuracy statements based on kappa coefficients derived from related samples. Moreover, it highlights that the proportion of correctly allocated cases, one of the most basic and widely used measures of accuracy, can be compared for both related and independent samples. Finally, a further feature apparent from the example is that it may sometimes be appropriate to conduct a one-sided test. This was illustrated with reference to the inclusion of prior information into a classification. Because the incorporation of prior information would be expected to increase classification accuracy, the direction of the alternative hypothesis used in the statistical testing can be specified. In the example presented, this had a major impact on the interpretation of the difference in accuracy between the classifications derived with and without prior information from the PNN.

In conclusion, therefore, when comparing thematic maps, researchers should be encouraged to statistically evaluate the significance of differences in map accuracy. There is no need to feel constrained to the use of comparison methods based on the kappa coefficient because many other approaches exist and may be preferred. The evaluation approach used should, however, appropriately recognize whether independent or related samples have been used in the derivation of the accuracy statements.

Acknowledgments

I am grateful to the Commission of the European Communities for the data sets which were acquired during the European AgriSAR campaign, Professor Dean McKenzie for helpful comments in relation to the comparison of related kappa coefficients and the three referees for their comments on the original manuscript.

References

Agresti, A., 1996. *An Introduction to Categorical Data Analysis*, Wiley, New York, N.Y., 312 p.

Bradley, J.V., 1968. *Distribution-Free Statistical Tests*, Prentice-Hall, Englewood Cliffs, New Jersey, 388 p.

Chan, J. C.-W., N. Laporte, and R.S. DeFries, 2003. Texture classification of logged forests in tropical Africa using machine learning algorithms, *International Journal of Remote Sensing*, 24: 1401–1407.

Chen, D.M., and D. Stow, 2002. The effect of training strategies on supervised classification at different spatial resolutions, *Photogrammetric Engineering & Remote Sensing*, 68:1155–1161.

Clark, W.A.V., and P.L. Hosking, 1986. *Statistical Methods for Geographers*, Wiley, New York, N.Y., 518 p.

Cohen, J., 1960. A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, 20:37–46.

Congalton, R.G., and K. Green, 1999. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, Lewis, Boca Raton, Florida, 137 p.

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*, 49:1671–1678.

Donner, A., M.M. Shoukri, N. Klar, and E. Bartfay, 2000. Testing the equality of two dependent kappa statistics, *Statistics in Medicine*, 19:393–387.

Fienberg, S.E., 1981. *The Analysis of Cross-classified Categorical Data, Second Edition*, MIT Press, Cambridge, Massachusetts, 161 p.

Fleiss, J.L., 1981. *Statistical Methods for Rates and Proportions, Second Edition*, Wiley, New York, N.Y., 352 p.

Foody, G.M., 2001. Thematic mapping from remotely sensed data with neural networks: MLP, RBF and PNN based approaches, *Journal of Geographical Systems*, 3:217–232.

_____, 2002. Status of land cover classification accuracy assessment, *Remote Sensing of Environment*, 80:185–201.

Freund, J.E., and F.J. Wiliams, 1959. *Modern Business Statistics*, Pitman, London, United Kingdom, 539 p.

Haack, B.N., E.K. Solomon, M.A. Bechdol, and N.D. Herold, 2002. Radar and optical data comparison/integration for urban delineation: A case study, *Photogrammetric Engineering & Remote Sensing*, 68:1289–1296.

Hayes, D.J., and S.A. Sader, 2001. Comparison of change-detection techniques for monitoring tropical forest clearing and vegetation regrowth in a time series, *Photogrammetric Engineering & Remote Sensing*, 67:1067–1075.

Hudson, W.D., and C.W. Ramm, 1987. Correct formulation of the kappa-coefficient of agreement, *Photgrammetric Engineering & Remote Sensing*, 53:421–422.

Janssen, L.L.F., and F.J.M. van der Wel, 1994. Accuracy assessment of satellite derived land-cover data: A review, *Photogrammetric Engineering & Remote Sensing*, 60:419–426.

McKenzie, D.P., A.J. Mackinnon, N. Peladeau, P. Onghena, P.C. Bruce, D.M. Clarke, S. Haarrigan, and P.D. McGorry, 1996. Comparing correlated kappas by resampling: Is one level of agreement significantly different from another?, *Journal of Psychiatric Research*, 30:483–492.

Neter, J., W. Wasserman, and G.A. Whitmore, 1993. *Applied Statistics, Fourth Edition*, Allyn and Bacon, Boston, Massachusetts, 989 p.

Rosenfield, G.H., and K. Fitzpatrick-Lins, 1986. A measure of agreement as a measure of thematic classification accuracy, *Photogrammetric Engineering & Remote Sensing*, 52:223–227.

Smits, P.C., S.G. Dellepiane, and R.A. Schowengerdt, 1999. Quality assessment of image classification algorithms for land-cover mapping: A review and a proposal for a cost-based approach, *International Journal of Remote Sensing*, 20:1461–1486.

Sohn, Y., and N.S. Rebello, 2002. Supervised and unsupervised spectral angle classifiers, *Photogrammetric Engineering & Remote Sensing*, 68:1271–1280.

Stehman, S.V., 1996. Estimating the kappa coefficient and its variance under stratified random sampling, *Photogrammetric Engineering & Remote Sensing*, 62:401–407.

_____, 1997. Selecting and interpreting measures of thematic classification accuracy, *Remote Sensing of Environment*, 62:77–89.

Thomas, I.L., and G. McK. Allcock, 1984. Determining the confidence level for a classification, *Photogrammetric Engineering & Remote Sensing*, 50:1491–1496.

Trodd, N.M., 1995. Uncertainty in land cover mapping for modelling land cover change, *Proceedings RSS95: Remote Sensing in Action* (Remote Sensing Society, Nottingham, 11–14 September, Southampton, United Kingdom) United Kingdom, pp. 1138–1145.

Zweig, M.H., and G. Campbell, 1993. Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool on clinical medicine, *Clinical Chemistry*, 39:561–577.

(Received 23 January 2003; accepted 26 March 2003; revised 23 May 2003)