Remote Sensing Brief

On the Compensation for Chance Agreement in Image Classification Accuracy Assessment

Giles M. Foody
Department of Geography, University College of Swansea, Singleton Park, Swansea SA2 8PP, United Kingdom

ABSTRACT: In the assessment of image classification accuracy with the kappa coefficient, the degree of chance agreement may be overestimated. Two kappa-like approaches which compensate more appropriately for the degree of chance agreement are discussed. Deriving these coefficients for a classification error matrix which had a percentage accuracy of 76.6 percent produced coefficients of approximately 0.69, higher than the kappa coefficient of 0.60. Because these alternative measures make a more appropriate compensation for chance agreement, they may be more suitable for the assessment of image classification accuracy than the kappa coefficient.

INTRODUCTION

A QUANTITATIVE MEASURE OF CLASSIFICATION ACCURACY is used typically to assess the quality of image classifications. Many such measures exist, ranging from general indicators of the proportion of cases allocated correctly by the classification to measures designed for specific applications (Hay, 1979; Aronoff, 1985; Story and Congalton, 1986). While no single measure is always appropriate for classification evaluation (Congalton, 1991), the kappa coefficient of agreement, $\hat{k}$, developed by Cohen (1960) and introduced to the remote sensing community in the early 1980s (Congalton and Mead, 1983; Congalton *et al.*, 1983) has become a widely used measure of classification accuracy. Its popularity arises primarily because *all* elements in the classification error matrix, and not just the main diagonal, contribute to its calculation and because it compensates for chance agreement (Rosenfield and Fitzpatrick-Lins, 1986; Campbell, 1987). The kappa coefficient has been considered generally as representing the proportion of agreement obtained after removing the proportion of agreement that could be expected to occur by chance. This brief article aims to show, however, that the calculation of the proportion of chance agreement for the calculation of the kappa coefficient may be overestimated, with a resultant under representation of classification accuracy.

THE KAPPA COEFFICIENT

The kappa coefficient of agreement may be calculated from Equation 1; i.e.,

$$\hat{k} = \frac{P_o - P_e}{1 - P_e} \quad (1)$$

where P_o is the observed proportion of agreement and P_e is the proportion of agreement that may be expected to occur by chance (Cohen, 1960; Rosenfield and Fitzpatrick-Lins, 1986). The latter is calculated from the row and column marginals of the classification error matrix from $\sum_{i=1}^{n} P_r(i)P_c(i)$, where n is the number of classes. The kappa coefficient lies typically on a scale between 0 and 1, where the latter indicates complete agreement, and is often multiplied by 100 to give a percentage measure of classification accuracy. Figure 1 illustrates the calculation of the kappa coefficient for one of the confusion matrices used by Congalton *et al.* (1983).

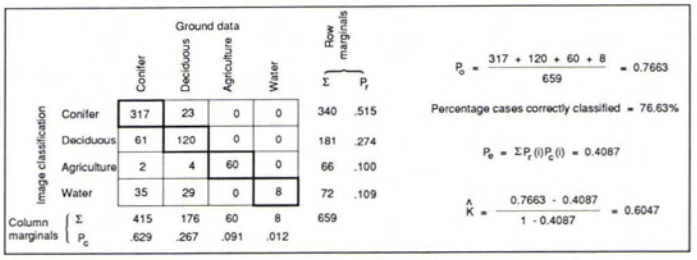

FIG. 1. Calculation of the kappa coefficient, $\hat{k}$, from a classification error matrix given by Congalton *et al.* (1983).

COMPENSATING FOR THE PROPORTION OF CHANCE AGREEMENT

The magnitude of P_e as defined above, however, includes actual agreement (Brennan and Prediger, 1981), or agreement for cause (Aickin, 1990), in addition to chance agreement. Consequently, the magnitude of the kappa coefficient calculated from Equation 1 will not reflect the proportion of agreement present in a classification less chance agreement only. Where the marginals are free (not fixed *a priori*), as is often the case in an image classification, then the marginal proportion for each class would be $1/n$. Consequently, the probability of chance agreement can be shown to be $1/n$, and an index of classification accuracy that could be considered as an alternative to the kappa coefficient could be defined as Equation 2 (Brennan and Prediger, 1981); i.e.,

$$k_n = \frac{P_o - 1/n}{1 - 1/n} \quad (2)$$

Calculating the k_n coefficient for the classification error matrix illustrated in Figure 1 gives $(0.7663 - 0.2500)/(1 - 0.2500) = 0.6884$, substantially higher than the kappa coefficient of 0.6047.

Another approach is to exclude the cases that are easy to classify which represent actual agreement, and so lie along the main diagonal of the classification error matrix, from the calculation of the proportion of chance agreement and utilize only the marginal probabilities for the cases that are hard to classify in the calculation of the proportion of chance agreement (Aickin,

PHOTOGRAMMETRIC ENGINEERING & REMOTE SENSING,
Vol. 58, No. 10, October 1992, pp. 1459–1460.

0099-1112/92/5810–1459$03.00/0

1990). The kappa-like statistic, α, proposed by Aickin (1990) can be derived from

$$\alpha = \frac{P_o - P_e}{1 - P_e}$$

with

$$p_r(i) = \frac{r(i)}{N(1 - \alpha + \alpha p_c(i)/P_e)}$$

and

$$p_c(i) = \frac{c(i)}{N(1 - \alpha + \alpha p_r(i)/P_e)}$$

where N is the total number of cases, $r(i)$ is the total in row i, and $c(i)$ is the total in column i of the classification error matrix. The statistic α is estimated by iteration until convergence is achieved. The kappa coefficient and observed row and column marginals are used as initial estimates for α, p_r, and p_c, respectively. Using this approach, after addition of a pseudo-count of 1 divided equally between all the elements of the classification error matrix to ensure convergence (Aickin, 1990), α was estimated to be 0.6925 for the data in Figure 1.

A number of other approaches may be used to provide measures of classification accuracy which effectively compensate more appropriately for chance agreement than the kappa coefficient. These include methods based on the quasi-independence concept (Goodman, 1975; Bergan, 1980) but these may be more difficult to interpret (Aickin, 1990).

SUMMARY AND CONCLUSIONS

Of the many approaches that may be used to assess image classification accuracy, the coefficient of agreement developed by Cohen (1960) has been used widely because it utilizes all the elements of the classification error matrix and attempts to remove the influence of chance agreement. The degree of chance agreement, however, may be overestimated in the calculation of the kappa coefficient because it is derived from the observed row and column marginals which include actual as well as chance agreement. This has the effect of lowering the magnitude of the kappa coefficient and so the apparent accuracy of the classification. Where users require a measure of classification accuracy which indicates the proportion of agreement present after the removal of chance agreement, then alternative kappa-like approaches such as those discussed by Brennan and Prediger (1981) and Aickin (1990) may be more appropriate than the kappa coefficient.

ACKNOWLEDGMENTS

I am grateful to Professor Paul Curran for his comments on a draft version of the manuscript.

REFERENCES

Aickin, M., 1990. Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa. *Biometrics,* 46:293-302.

Aronoff, S., 1985. The minimum accuracy value as an index of classification accuracy. *Photogrammetric Engineering & Remote Sensing,* 51:99-111.

Bergan, J. R., 1980. Measuring observer agreement using the quasi-independence concept. *Journal of Educational Measurement,* 17:59-69.

Brennan, R. L., and D.J. Prediger, 1981. Coefficient kappa: some uses, misuses, and alternatives. *Educational and Psychological Measurement,* 41:687-699.

Campbell, J. B., 1987. *Introduction to Remote Sensing,* Guildford Press, New York.

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

Congalton, R. G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment,* 37:35-46.

Congalton, R. G., and R. A. Mead, 1983. A quantitative method to test for consistency and correctness in photo-interpretation. *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.

Goodman, L. A., 1975. A new model for scaling response patterns: an application of the quasi-independence concept. *Journal of the American Statistical Association,* 70:755-768.

Hay, A. M., 1979. Sampling designs to test land-use map accuracy. *Photogrammetric Engineering & Remote Sensing,* 45:529-533.

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.

Story, M., and R. G. Congalton, 1986. Accuracy assessment: A user's perspective. *Photogrammetric Engineering & Remote Sensing,* 52:397-399.

(Received 21 October 1991; accepted 22 April 1992)