ON CHANCE-ADJUSTED MEASURES FOR ACCURACY ASSESSMENT IN REMOTE SENSING IMAGE CLASSIFICATION

Shiguo Jiang
Department of Geography
The Ohio State University
1036 Derby Hall, 154 N Oval Mall
Columbus, OH 43210
jiang.152@osu.edu

Desheng Liu
Department of Geography and Department of Statistics
The Ohio State University
1036 Derby Hall, 154 N Oval Mall
Columbus, OH 43210
liu.738@osu.edu

ABSTRACT

The underlying rationale and practical utility of chance-adjusted indices (e.g., kappa, tau) as accuracy measures in image classification have been under criticism for a long time despite the fact that they are near universally used. It has been suggested that the degree of chance agreement may be overestimated, or it makes no sense to use kappa or tau for their declared objectives due to the inconsistency of the chance definition. On the contrary, user's accuracy, producer's accuracy, and overall accuracy should be recommended because they are directly interpretable as probabilities of correct classification. Besides the continuing criticism in remote sensing literature, much more discussions can be found in psychology and sociology literature where kappa originated. In this paper, we give a review on literature of the chance-adjusted measures, specifically kappa-like measures. We focus our discussion on whether those measures are theoretically sound and practically interpretable. We re-evaluate the usefulness of kappa-like measures and give our recommendation of proper accuracy measures for accuracy assessment.

KEYWORDS: Accuracy assessment, chance-adjusted measure, kappa, remote sensing, image classification

INTRODUCTION

Accuracy assessment plays an important role in remote sensing image classification. It is important to know the quality of the classification maps before we conduct further analysis. A number of indices have been proposed to measure the accuracy of classification maps, among which overall accuracy (OA), producer's accuracy (PA), user's accuracy (UA), and kappa (κ) are mostly used.

Overall accuracy, also called overall agreement, raw accuracy, or proportion of pixels correctly classified, is the proportion of pixels whose class labels agree with the ground reference. It is suggested that overall accuracy includes chance agreement indicated by the row and column totals in the error matrix and the expected chance highly depends on the number of classes in the image classification (Cohen 1960). Therefore, it is declared that overall accuracies from different image classifications are not suitable for comparison when the number of classes is different.

Chance-adjusted measures like kappa are proposed to overcome the comparability issue of raw measures such as overall accuracy. Chance-adjusted agreement is measured by removing the chance agreement and is supposed to provide a better index for accuracy assessment. In this paper, we adopted the view of Stehman (1997) that “chance-adjusted” is a better terminology than “chance-corrected”.

Chance-adjusted measures have been under continuing criticism since 1980s. Chance-adjusted measures were first developed in sociology and psychology and then introduced to remote sensing community. Criticisms follow the same path of knowledge transfer. Main works criticizing chance-adjusted measures in remote sensing include Foody (1992, 2008), Ma and Redmond (1995), Stehman (1997; 1999), Pontius (2000), Liu, Frazier, and Kumar (2007), and Stehman

and Foody (2009). Despite the continuing efforts of these scholars, the criticisms are to a large extent ignored by most studies. Researchers often feel that it is obligated to report kappa in their research.

This paper aims to provide a review and re-evaluation of the chance-adjusted measures and related criticisms. Recommendations on the use of chance-adjusted measures are also provided. Among all the chance-adjusted measures, kappa is the most widely used one. Therefore, the paper mainly focuses on kappa-like measures.

DEVELOPMENT OF KAPPA-LIKE MEASURES

Kappa-like Measures

Kappa-like measures are a collection of chance-adjusted indices used to account for the accuracy of image classification (interpretation) that can be attributable to random chance. Here we mainly review those measures that are familiar to the remote sensing community, including kappa, weighted kappa, conditional kappa, and tau (**Table 1**). Besides kappa-like measures, other more complex chance-adjusted measures have also been proposed, such as Aickin (1990)'s α, Andres and Marzo (2004)'s delta, etc. However, those complex measures are of little use due to the difficulty of interpretation.

Table 1. Main chance-adjusted measures shared by different disciplines

Measures	Literature in sociology and psychology	Literature in remote sensing
Kappa	Cohen (1960)	Congalton (1980, 1981), Chrisman (1980), Congalton, Oderwald, and Mead (1983)
Conditional kappa	Coleman (1966), Light (1971)	Rosenfield and Fitzpatrick (1986)
Weighted kappa	Coleman (1968)	Rosenfield and Fitzpatrick (1986)
Tau	Klecka (1980)	Ma and Redmond (1995)

The above four kappa-like measures are defined based on error matrix (also called confusion matrix) where the results of two raters or judges are compared. In the case of image analysis in remote sensing, the results of image classification are compared to the ground reference. Suppose N pixels are considered, and they are classified into n classes in an image classification process (**Table 2**).

Table 2. Error matrix

	Ground reference					Row total	User's accuracy
Image classification	Class	1	2	...	n		
	1	x_{11}	x_{12}	...	x_{1n}	x_{1+}	x_{11}/x_{1+}
	2	x_{21}	x_{22}	...	x_{2n}	x_{2+}	x_{22}/x_{2+}
	...	...	...	...	...	...	...
	n	x_{n1}	x_{n2}	...	x_{nn}	x_{n+}	x_{nn}/x_{n+}
Column total		x_{+1}	x_{+2}	...	x_{+n}	N	
Producer's accuracy		x_{11}/x_{+1}	x_{22}/x_{+2}	...	x_{nn}/x_{+n}		

Let x_{ij} be the number of pixels that are classified as class i, but are actually class j in the ground reference. The column totals x_{+k} and row totals x_{k+} are called marginals. The four kappa-like measures have the universal form as

$$\text{Index} = \frac{P_o - P_c}{\max(P_o) - P_c}, \tag{1}$$

where P_o is the overall accuracy, max(P_o) is the maximum possible overall accuracy that can be observed, and P_c is the proportion of pixels for which agreement is expected by random chance. Equation (1) has the same form as the index β proposed by Brennan and Prediger (1981). For all kappa-like indices, P_o is defined as

$$P_o = \frac{1}{N}\sum_{i=1}^{n} x_{ii}, \qquad N = \sum_{i=1}^{n} x_{i+} = \sum_{i=1}^{n} x_{+i}. \tag{2}$$

The definition of max(P_o) and P_c varies with the index defined.

Kappa. The popular kappa index was first introduced as a new technique by the statistician and psychologist Jacob Cohen in his seminal paper published in the journal *Education and Psychological Measurement* in 1960 (Cohen 1960). Based on Türk (2002)'s review, kappa was introduced to the remote sensing community in early 1980s by Russell G. Congalton and his co-workers (Congalton 1980, 1981; Congalton, Oderwald, and Mead 1983) in USA and by Chrisman (1980) in Britain. Cohen defined max (P_o) as 1, and chance agreement P_c as

$$P_c = \frac{1}{N^2}\sum_{i=1}^{n} x_{i+}x_{+i}. \tag{3}$$

Cohen's kappa is then defined as

$$\kappa = \frac{P_o - P_c}{1 - P_c} = \frac{N\sum_{i=1}^{n} x_{ii} - \sum_{i=1}^{n} x_{i+}x_{+i}}{N^2 - \sum_{i=1}^{n} x_{i+}x_{+i}}. \tag{4}$$

The limits (i.e., maximum and minimum) and sampling characteristics (i.e., variance/standard error and confidence intervals) of kappa were also discussed by Cohen (1960) and a correction to the variance of kappa was given later by Cohen and his co-workers (Fleiss, Cohen, and Everitt 1969).

Kappa is declared to have two advantages over raw accuracy:

- Kappa considers all the cells of an error matrix and thus incorporates more information (Rosenfield and Fitzpatrick-lins 1986; Fung and Ledrew 1988; Dicks and Lo 1990; Jansen and van der Wel 1994);
- Kappa is suitable for comparison between different error matrices because it removes chance agreement (Congalton, Oderwald, and Mead 1983; Congalton 1991).

The estimated kappa coefficients of two maps, $\hat{\kappa}_1$ and $\hat{\kappa}_2$, are usually compared to examine the significance of difference in image classification accuracy. The significance is defined as

$$z = \frac{\hat{\kappa}_1 - \hat{\kappa}_2}{\sqrt{\hat{\sigma}_1 + \hat{\sigma}_2}}, \tag{5}$$

where $\hat{\sigma}_1$ and $\hat{\sigma}_2$ represents the estimated variances of the kappa coefficients $\hat{\kappa}_1$ and $\hat{\kappa}_2$ respectively.

Weighted kappa. Weighted kappa was proposed to consider partial agreement, errors of varying importance, or agreement of ordinal data (Cohen 1968; Fleiss, Cohen, and Everitt 1969). For example, it may be worse to classify an agriculture area as bare land than to classify it as grass land. Weighted kappa was introduced to remote sensing by Rosenfield and Fitzpatrick-lins (1986). Assigning weight w_{ij} to the i, jth cell in an error matrix, the weighted overall agreement P_o^*, and chance agreement P_c^* are defined as

$$P_o^* = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}x_{ij}, \qquad P_c^* = \frac{1}{N^2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}x_{i+}x_{+j}. \tag{6}$$

Thus weighted kappa is given by

$$\kappa_w = \frac{P_o^* - P_c^*}{1 - P_c^*}. \tag{7}$$

Conditional kappa. Conditional kappa was proposed to test the individual category agreement (Coleman 1966; Light 1971) and was introduced to remote sensing by Rosenfield and Fitzpatrick-lins (1986). P_o, P_c, and the maximum agreement available for the ith category are defined as

$$P_{oi} = \frac{x_{ii}}{N}, \qquad P_{ci} = \frac{x_{i+}x_{+i}}{N^2}, \qquad P_{\max i} = \frac{x_{i+}}{N}. \tag{8}$$

Thus conditional kappa for the ith category is given by

$$\kappa_i = \frac{P_{oi} - P_{ci}}{P_{\max i} - P_{ci}} = \frac{Nx_{ii} - x_{i+}x_{+i}}{Nx_{i+} - x_{i+}x_{+i}}. \tag{9}$$

Tau. Tau measures the agreement by comparing the classification with a random assignment of pixels to classes (Klecka 1980). It can be regarded as a response to the criticisms of kappa (more discussion to follow in the next section). Tau was introduced to remote sensing by Ma and Redmond (1995). Tau is defined as

$$T = \frac{P_o - P_c}{1 - P_c}, \tag{10}$$

where P_o is as defined before (referring to equation (2)), and P_c is defined as

$$P_c = \sum_{i=1}^{n} \frac{x_{+i}}{N} \cdot \frac{x_i}{N} = \frac{1}{N^2} \sum_{i=1}^{n} x_{+i} x_i. \tag{11}$$

In equation (11), x_i /N is user-specified, the *a priori* probability of class membership, and x_i is proportional to x_{+i}. When the *a priori* probabilities of class membership are equal for a classification, $x_i = N/n$, thus

$$T_n = \frac{P_o - 1/n}{1 - 1/n}. \tag{12}$$

Equation (12) is a special case of tau, and it is the same as the modified kappa statistic, κ_n, proposed by Brennan and Prediger (1981) and Foody (1992).

CRITICISMS

Criticisms on chance-adjusted measures are mainly about the declared advantages and properties of kappa as it is the most widely used one, mainly on the declared advantages and properties. There are more criticisms of kappa in other fields than in image classification. However, some criticisms are not relevant to image classification due to the way that kappa is used in different disciplines. For example, the way kappa used in image classification is different from that in sociology and psychology. In the latter, the marginals (i.e., column totals and row totals of the error matrix in **Table 2**) can be either fixed or free. When both marginals are fixed, column total x_{+k} are deemed to equal to the corresponding row total x_{k+}. In accuracy assessment of image classification, the column totals are from ground reference and are thus fixed. The row totals are from image classification, and are free. Therefore, x_{+k} does not necessary equal x_{k+}. Therefore, many criticisms on kappa as to the marginal issues are not relevant to image classification (Brennan and Prediger 1981; Cicchetti and Feinstein 1990; Feinstein and Cicchetti 1990; von Eye and von Eye 2008). In the following, we only review those

criticisms relevant to image classification.

Kappa does not Consider All the Cells in Error Matrix

As stated in the previous section, one of kappa's declared advantages is its use of information of all the cells from the error matrix. Criticisms of this advantage are straightforward and leave no confusion (Stehman 1997; Nishii and Tanaka 1999). Based on the definition of kappa, the chance term P_c is defined based on the marginals, i.e., row totals and column totals; therefore kappa does incorporate some off-diagonal information in the error matrix (Stehman 1997). However, the total of cells cannot reflect the detailed information of all cells. Kappa does not use all the cells since different internal configurations of the error matrix can result in the same marginals (Stehman 1997). For example, if we swap the cells of the upper triangle with the lower triangle in **Table 2**, we would get the same kappa value.

Controversies in the Definition of Chance Agreement

Criticisms of the second advantage of kappa are critical and need more attention. It is found that the introduction of chance agreement is problematic. The definition of chance in kappa is based on the assumption that row totals and column totals are independent and randomly assigned. Brennan and Prediger (1981) provide one of the first critical evaluations of kappa, whose relevant points to image classification are summarized here. Brennan and Prediger (1981) argue that the chance agreement should be defined as $P_c = 1/n$ based on the independent and random assumption. This argument is supported by Foody (1992) who argues that chance agreement is overestimated in kappa resulting in an underestimation of classification accuracy. Later, Ma and Redmond (1995) find that Brennan and Prediger (1981)'s κ_n is a special case of tau (Equation (12)). The main difference between kappa and tau coefficient is that kappa uses the marginal proportions of the classified map to define chance agreement, whereas tau uses marginal proportions specified *prior* to image classification. In other words, kappa is based on the *a posteriori* probabilities of class membership, whereas tau is based on the *a priori* probabilities.

Uebersax (1987) may be the first to notice the contradiction in the definition of chance agreement using the *a posteriori* probabilities. The term P_c represents chance agreement accountable to the null hypothesis of randomly assigning class membership. There would be no interpretation problem of kappa when the null hypothesis is true. In other words, when both the reference data and the classification map are randomly assigned with class members, P_c could be calculated as defined. Kappa would equal zero in this case. However, it would be problematic to interpret kappa when the null hypothesis is not true. This second situation is actually quite common in practice. As neither the reference data nor classification map is randomly assigned, chance agreement P_c cannot be calculated as defined. Therefore, it makes no sense to calculate kappa based on P_c.

Agresti (1996) further discusses the controversy and circularity in reasoning of kappa. Based on the definition of chance agreement in kappa, the marginal proportions of the row totals are fixed before the classification. However, the marginal proportions used in calculating chance agreement P_c are the result of the classification, not fixed marginal proportions. Therefore, the calculation of P_c violates its definition, which leads to the problems of calculating and interpreting kappa.

While acknowledging the issues of using the *a posteriori* probabilities to calculate the chance agreement as defined, Stehman (1997) further argues that it is also problematic to define chance agreement using the *a priori* probabilities as in tau index. Stehman (1997) identifies two issues of tau. First, the dependence of tau on the *a priori* probabilities causes confusion. Two maps with the same error matrices may have different tau coefficients because the *a priori* probabilities (say β_k) are different for the two maps. Second, it is difficult to interpret tau if the map marginal proportions are not forced to match β_k in the classification process. This is usually the case in image classification as the marginal proportions of the rows are usually free.

Another criticism is that using the *a priori* probabilities introduces uncertainty to the calculation of tau (Stehman 1997; Liu, Frazier, and Kumar 2007). For example, there may not be available data to estimate the *a priori* probabilities. The determination of the *a priori* probabilities may also be subjective.

Conditional kappa and weighted kappa also suffer the same criticisms applied to kappa and tau as they inherit the same framework of kappa (Liu, Frazier, and Kumar 2007). Moreover, the definition of weight introduces further uncertainties to weighted kappa.

Misuse of Kappa

The widely used guidelines to interpret the magnitude of kappa as strength of agreement are problematic (Manel,

Williams, and Ormerod 2001; Di Eugenio and Glass 2004; von Eye 2008; Foody 2008). **Table 3** gives two guidelines which are based on the interpretation of kappa as a measure of degree of agreement. However, kappa is defined as the agreement beyond chance, not agreement per se. Rules of thumb as in **Table 3** can be misleading. Kappa is often mis-interpreted as a measure of agreement per se in the literature. For example, Warren et al. (2002) only reported kappa when comparing the accuracy of image classification from Landsat imagery and SAR imagery (Table 11.2 and 11.3 on p.176). They stated that (p. 175; Italics added), "While the *accuracy* derived from the SAR data was fairly low, the *accuracy* for black spruce was higher than observed in the Landsat TM-based classification."

Table 3. Strength of agreement based on kappa

(A) Landis and Koch (1977)'s scheme				(B) Fleiss (1981)'s scheme	
Kappa	Agreement	Kappa	Agreement	Kappa	Agreement
<0.00	Poor	0.41-0.60	Moderate	<0.40	Poor
0.00-0.20	Slight	0.61-0.80	Substantial	0.40-0.75	Good
0.21-0.40	Fair	0.81-1.00	Almost Perfect	0.75-1	Excellent

Note: The inconsistency of the two schemes itself also indicates the arbitrary and problematic of the scale division.

There are fundamental concerns associated with the accuracy comparison based on Equation (5) (Foody 2004). Equation (5) can only be used when independent samples are applied in calculating kappa (Cohen 1960). However, this assumption of independence is not always satisfied. Quite often the samples used to calculate kappa are related. For example, kappa coefficients to be compared are calculated based on the same sample sites (Congalton, Oderwald, and Mead 1983; Haack et al. 2002; Sohn and Rebello 2002). A detailed discussion of this issue can be found in Foody (2004), with solutions of this issue proposed.

DISCUSSION AND CONCLUSIONS

The main points in previous section are summarized as follows (in the case of kappa-like measures):

- Kappa does consider all the cells in the error matrix as declared.
- The definition of chance agreement bears fatal controversy and is not applicable in practice.
- Kappa is quite often misinterpreted in literature as a measure of agreement per se, not agreement beyond chance as defined.
- The comparisons of kappa coefficients from different classifiers are often wrong due to the use of related samples.

Given the review of chance-adjusted measures and related criticisms, we may ask the following critical question: Are chance-adjusted measures necessary? Some researchers have suggested that chance-corrected measures are not necessary. For example, Stehman (1997; 1999) suggests that it makes no sense to use kappa and tau for their declared objectives due to the inconsistency of the chance definition. On the contrary, overall accuracy (P_o,), producer's accuracy (PA) and user's accuracy (UA) should be recommended as they are directly interpretable as probabilities of correct classification. Türk (2002) argues that chance adjustment is completely unnecessary and the traditional practice of reporting kappa should be admonished. It is suggested that the correct classification due to chance is a windfall gain; it is not necessary for the users or produers of the maps to worry about.

Some declared characteristics/advantages of chance-adjusted measures actually also apply equally to other measures of accuracy (Foody 2008). For example, variance term can be derived for other widely used accuracy measures such as overall accuracy (Foody 2004). The statistical significance of differences in classification accuracy can then be evaluated based on this newly derived variance term.

The equivalence between overall accuracy and chance-adjusted measure have also been found if the inconsistency of chance agreement is ignored. For example, Stehman (1997) argues that T_n (Equation (12)) is actually a linear scaling of overall accuracy (P_o). Therefore, T_n rescales the magnitude of the difference between the P_o values without changing the ordering or ranking. Similar relationship has also been found between overall accuracy and kappa. In an review of papers published in the journal *Photogrammetric Engineering and Remote Sensing* from 1989 until 2003, Wilkinson (2005) found that there is strong linear correlation between kappa and overall accuracy. This high correlation has also been

confirmed by Liu, Frazier, and Kumar (2007) in their extensive study of published error matrices. This means that overall accuracy and kappa provide much the same information.

In conclusion, we do not recommend the use of chance-adjusted measures, considering to three fatal unfavorable characteristics: (1) inconsistency in the definition of chance agreement; (2) misleading interpretation and misuse in practice; (3) no more informative information compared to other measures such as overall accuracy. We do recommend that overall accuracy, producer's accuracy, and user's accuracy be provided along with the original error matrix from which these three accuracy measures are calculated. When these four items are presented together, the readers will get all the necessary information as to the global image accuracy.

ACKNOWLEDGEMENT

The research was supported by The Climate, Water, and Carbon (CWC) Initiative at The Ohio State University.

REFERENCES

Agresti, A. 1996. *An Introduction to Categorical Data Analysis*. New York: Wiley.

Aickin, M. 1990. Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa. *Biometrics* 46 (2):293-302.

Andres, A. M., and P. F. Marzo. 2004. Delta: A new measure of agreement between two raters. *British Journal of Mathematical & Statistical Psychology* 57:1-19.

Brennan, R. L., and D. J. Prediger. 1981. Coefficient kappa: some uses, misuses, and alternatives. *Educational and Psychological Measurement* 41 (3):687-699.

Chrisman, N. R. 1980. (September 10). Assessing Landsat accuracy: a geographic application of misclassification analysis. In *Second Colloquim on Quantitative and Theoretical Geography*. Trinity Hall, Cambridge, England. (cited in James B. Campbell's Introduction to Remote Sensing, 2nd Ed., The Guilford Press, Chap. 13).

Cicchetti, D. V., and A. R. Feinstein. 1990. High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology* 43 (6):551-558.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20 (1):37-46.

———. 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 70 (4):213-220.

Coleman, J. S. 1966. Measuring Concordance in Attitudes, 43p. Baltimore, MD: Department of Social Relations, Johns Hopkins University.

Congalton, R. G. 1980. (March 11).Statistical techniques for analysis of Landsat classification accuracy. In *Meeting of the American Society of Photogrammetry*. St. Louis, MO (cited by Rosenfield and Fitzpatrick-Linz, 1986).

———. 1981. The use of discrete multivariate analysis for the assessment of Landsat classification accuracy. Mater thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA.

———. 1991. A review of assesing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment* 37 (1):35-46.

Congalton, R. G., R. G. Oderwald, and R. A. Mead. 1983. Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques. *Photogrammetric Engineering and Remote Sensing* 49 (12):1671-1678.

Di Eugenio, B., and M. Glass. 2004. The kappa statistic: A second look. *Computational Linguistics* 30 (1):95-101.

Dicks, S. E., and T. H. C. Lo. 1990. Evaluation of thematic map accuracy in a land-use and land-cover mapping program. *Photogrammetric Engineering and Remote Sensing* 56 (9):1247-1252.

Feinstein, A. R., and D. V. Cicchetti. 1990. High agreement but low Kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology* 43 (6):543-549.

Fleiss, J. L. 1981. *Statistical methods for rates and proportions*. 2nd ed. New York: Wiley.

Fleiss, J. L., J. Cohen, and B. S. Everitt. 1969. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin* 72 (5):323-&.

Foody, G. M. 1992. On the compensation for chance agreement in image classification accuracy assessment.

Photogrammetric Engineering and Remote Sensing 58 (10):1459-1460.

———. 2004. Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy. *Photogrammetric Engineering and Remote Sensing* 70 (5):627-633.

———. 2008. Harshness in image classification accuracy assessment. *International Journal of Remote Sensing* 29 (11):3137-3158.

Fung, T., and E. Ledrew. 1988. The determination of optimal threshold levels for change detection using various accuracy indices. *Photogrammetric Engineering and Remote Sensing* 54 (10):1449-1454.

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 and Remote Sensing* 68 (12):1289-1296.

Jansen, L. L. F., and F. J. M. van der Wel. 1994. Accuracy assessment of satellite derived land-cover data: a review. *Photogrammetric Engineering and Remote Sensing* 60 (4):419-426.

Klecka, W. R. 1980. *Discriminant Analysis*. Beverly Hills, CA: SAGE Publication, Inc.

Landis, J. R., and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33 (1):159-174.

Light, R. J. 1971. Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychological Bulletin* 76 (5):365-377.

Liu, C. R., P. Frazier, and L. Kumar. 2007. Comparative assessment of the measures of thematic classification accuracy. *Remote Sensing of Environment* 107 (4):606-616.

Ma, Z. K., and R. L. Redmond. 1995. Tau coefficients for accuracy assessment of classification of remote sensing data. *Photogrammetric Engineering and Remote Sensing* 61 (4):435-439.

Manel, S., H. C. Williams, and S. J. Ormerod. 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology* 38 (5):921-931.

Nishii, R., and S. Tanaka. 1999. Accuracy and inaccuracy assessments in land-cover classification. *Ieee Transactions on Geoscience and Remote Sensing* 37 (1):491-498.

Pontius, R. G. 2000. Quantification error versus location error in comparison of categorical maps. *Photogrammetric Engineering and Remote Sensing* 66 (8):1011-1016.

Rosenfield, G. H., and K. Fitzpatrick-lins. 1986. A coefficient of agreement as a measure of thematic classification accuracy. *Photogrammetric Engineering and Remote Sensing* 52 (2):223-227.

Sohn, Y., and N. S. Rebello. 2002. Supervised and unsupervised spectral angle classifiers. *Photogrammetric Engineering and Remote Sensing* 68 (12):1271-1282.

Stehman, S. V. 1997. Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment* 62 (1):77-89.

———. 1999. Comparing thematic maps based on map value. *International Journal of Remote Sensing* 20 (12):2347-2366.

Stehman, S. V., and G. M. Foody. 2009. Accuracy Assessment. In *The Sage handbook of remote sensing*, eds. T. A. Warner, M. D. Nellis and G. M. Foody, 297-309. Los Angeles, CA: Sage.

Türk, G. 2002. Chance correction and map evaluation. *Remote Sensing of Environment* 82 (1):1-3.

Uebersax, J. S. 1987. Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin* 101 (1):140-146.

von Eye, A. 2008. What do we know about Cohen's kappa? A review and discussion. In *Configural Frequency analysis (CFA) and other nonparametric statistical methods: Gustav A. Lienert Memorial Issue*, eds. M. Stemmler, E. Lautsch and D. Martinke, p. 29-39. Lengerich: Pabst.

von Eye, A., and M. von Eye. 2008. On the Marginal Dependency of Cohen's kappa. *European Psychologist* 13 (4):305-315.

Warren, A. J., M. J. Collins, E. A. Johnson, and P. F. Ehlers. 2002. Managing uncertainty in a geospatial model of biodiversity. In *Uncertainty in remote sensing and GIS*, eds. G. Foody and P. Atkinson, 167-185. Chichester, England: John Wiley & Sons Ltd.

Wilkinson, G. G. 2005. Results and implications of a study of fifteen years of satellite image classification experiments. *Ieee Transactions on Geoscience and Remote Sensing* 43 (3):433-440.