Selecting and interpreting measures of thematic classification accuracy

Article *in* Remote Sensing of Environment · October 1997

DOI: 10.1016/S0034-4257(97)00083-7

1 author:

Stephen V. Stehman
State University of New York College of Environmental Science and Forestry
225 PUBLICATIONS **28,794** CITATIONS

SEE PROFILE

ELSEVIER

Selecting and Interpreting Measures of Thematic Classification Accuracy

Stephen V. Stehman

An error matrix is frequently employed to organize and display information used to assess the thematic accuracy of a land-cover map, and numerous accuracy measures have been proposed for summarizing the information contained in this error matrix. No one measure is universally best for all accuracy assessment objectives, and different accuracy measures may lead to conflicting conclusions because the measures do not represent accuracy in the same way. Choosing appropriate accuracy measures that address objectives of the mapping project is critical. Characteristics of some commonly used accuracy measures are described, and relationships among these measures are provided to aid the user in choosing an appropriate measure. Accuracy measures that are directly interpretable as probabilities of encountering certain types of misclassification errors or correct classifications should be selected in preference to measures not interpretable as such. User's and producer's accuracy and the overall proportion of area correctly classified are examples of accuracy measures possessing the desired probabilistic interpretation. The kappa coefficient of agreement does not possess such a probabilistic interpretation because of the adjustment for hypothetical chance agreement incorporated into this measure, and the strong dependence of kappa on the marginal proportions of the error matrix makes the utility of kappa for comparisons suspect. Normalizing an error matrix results in estimates that are not consistent for accuracy parameters of the map being assessed, so that this procedure is generally not warranted for most applications. ©*Elsevier Science Inc., 1997*

SUNY College of Environmental Science and Forestry, Syracuse, New York

Address correspondence to S. V. Stehman, SUNY Coll. of Environmental Science & Forestry, 320 Bray Hall, Syracuse, NY 13210.
Received 19 December 1996; revised 30 April 1997.

INTRODUCTION

Accuracy assessment of land-cover maps constructed from remotely sensed data contributes important data quality information to map users. Accuracy assessment is usually conducted by selecting a sample of reference locations, and comparing the classifications at these reference locations to the classifications provided by the land-cover map. The reference sample should be selected independently of data used for training and/or developing the classification procedure. The reference sample data are then summarized in an error matrix (Congalton et al., 1983; Story and Congalton, 1986), and various accuracy statistics are computed from this error matrix.

A variety of measures have been suggested for describing the accuracy of land-cover classifications. The overall proportion of area, pixels or polygons classified correctly for the entire map, various forms of kappa (κ) coefficients of agreement, the τ coefficient, user's and producer's accuracy, and conditional κ are commonly used accuracy measures. No consensus has been reached on which measures are appropriate for a given objective of accuracy assessment, although the kappa statistic seems to be generally favored. Rosenfield and Fitzpatrick-Lins (1986, p. 226) recommended "that coefficients of Kappa and conditional Kappa be adopted by the remote sensing community as a measure of accuracy for thematic classification as a whole, and for the individual categories." Fitzgerald and Lees (1994, p. 368) proposed "that the Kappa test statistic be used in preference to the overall accuracy as a means of testing classification accuracy based on error matrices." Fung and LeDrew (1988, p. 1453) concluded that "accuracy indices based on the producer's accuracy and overall accuracy may tend to be biased towards the category with a large number of samples," and recommended that κ be used because "all cells of the error matrix are considered." But Foody (1992) suggested that the usual κ overestimated chance agreement and recommended a modified form of

REMOTE SENS. ENVIRON. 62:77–89 (1997)

655 Avenue of the Americas, New York, NY 10010

0034-4257/97/$17.00
PII S0034-4257(97)00083-7

κ. Ma and Redmond (1995, p. 435) showed that this modified κ could be viewed as another type of agreement measure called τ, and then stated that τ "is a better measure of classification accuracy for use with remote sensing data than either Kappa or percentage agreement." Lark (1995) mentions the use of κ in accuracy assessment, but does not describe an application in which κ is the parameter of choice among the several examples presented.

Each accuracy measure provides a different summary of the information contained in an error matrix, and each may be applicable for a particular user in a given project. Lark (1995) describes specific circumstances in which one accuracy measure may be more relevant than others for a particular objective. Congalton (1991) suggested calculating various accuracy measures, and if the interpretations differed, evaluating the nature of the differences. Given that different accuracy measures are appropriate for different objectives, it is important to understand the characteristics of each measure. The objective of this article is to describe properties of and relationships among these accuracy measures and to illustrate why they differ in certain circumstances. Supplied with this information, a user can better decide which accuracy measure is appropriate for a given application and objective.

Uses of Map Accuracy Measures

Because an error matrix is such an effective descriptive tool for organizing and presenting accuracy assessment information, the error matrix should be reported whenever feasible. In addition to the valuable descriptive role of the full error matrix, various map accuracy measures will be of interest to summarize the error matrix information. Uses of these summary measures may be grouped into two broad application classes, reporting the accuracy of a final map product, and comparing maps.

In the first class of applications, a user has available a final product, land-cover map and the accuracy assessment objective is to provide a description of classification error (e.g., Congalton et al., 1993; Dicks and Lo, 1990; Fiorella and Ripple, 1993; Lauver and Whistler, 1993; Lawrence et al., 1996; Vujakovic, 1987). In the second class of applications, the map producers may still be in the process of creating a map for the region of interest. In this application, the identity and number of land-cover classes is often the same, and the objective is to determine which map constructed from the candidate imagery dates, classification algorithms, or other available options results in the highest accuracy. For example, Gong and Howarth (1990) evaluated several factors influencing classification accuracy, Treitz et al. (1992) compared the accuracies of maps derived from different methods for classifying training data, and Stenback and Congalton (1990) compared accuracies of different TM band combinations. The first class of applications requires selection of the most appropriate descriptor of map accuracy, whereas the map comparison application requires an appropriate descriptor, an ability to rank the maps, and a measure of the magnitude of the difference in accuracy.

Table 1. Population Error Matrix with p_{ij} Representing the Proportion of Area in Mapped Land-Cover Class i and Reference Land-Cover Class j

		Reference				*Row*
		1	*2*	...	q	*Proportion*
Classified	1	p_{11}	p_{12}	...	p_{1q}	p_{1+}
	2	p_{21}	p_{22}	...	p_{2q}	p_{2+}
	:	:	:	...	:	:
	q	p_{q1}	p_{q2}	...	p_{qq}	p_{q+}
Col. proportion		p_{+1}	p_{+2}	...	p_{+q}	

Sometimes comparisons are needed for maps of different regions and/or different land-cover classification schemes. Such applications present perhaps the most difficult task because of the confounding of accuracy with regional and land-cover scheme differences. For example, even if the same region is represented by the two maps but the land-cover classification schemes differ, the user likely has different objectives motivating the two schemes, and a direct comparison based only on a map accuracy measure may not capture this difference in objectives.

Notation and Description of Accuracy Parameters

The accuracy assessment measures can be defined as parameters of a population error matrix. That is, given complete reference data for the study region, both the reference and image classifications for all areas on the map are available. A population error matrix (Table 1) could be constructed from this census, where p_{ij} is the probability that a randomly selected area is classified as category i by the image and as category j by the reference data. Summary measures computed from this error matrix are then population parameters. Recognizing that these parameters are characteristics of a real population is a useful device when interpreting the various accuracy measures. In practice, a sampling design is implemented, and the sample data obtained are used to estimate the population error matrix and associated parameters. However, to examine the various accuracy measures, it is simpler to operate at the population level and forego the specifics of how to estimate these parameters for a particular sampling design.

The accuracy parameters that will be discussed are listed below. In the Table 1 notation, lower case p will be used to denote characteristics of the population error matrix such as the individual cell probabilities (p_{ij}) and row and column marginal proportions (p_{i+} and p_{+j}, re-

spectively). Upper case P will be used to denote summary parameters such as the overall proportion of area correctly classified (P_c), and user's (P_{Ui}) and producer's (P_{Aj}) accuracy.

1. Overall proportion of area correctly classified,

$$P_c = \sum_{k=1}^{q} p_{kk}, \tag{1}$$

where q=number of land-cover categories.

2. Kappa,

$$\kappa = \frac{P_c - \sum_{k=1}^{q} p_{k+}p_{+k}}{1 - \sum_{k=1}^{q} p_{k+}p_{+k}}, \tag{2}$$

where $p_{k+} = \sum_{j=1}^{q} p_{kj}$, and $p_{+k} = \sum_{i=1}^{q} p_{ik}$.

3. Kappa with random chance agreement as defined by Foody (1992),

$$\kappa_e = \frac{P_c - 1/q}{1 - 1/q}, \tag{3}$$

the subscript e used to indicate that each land-cover class is equally probable under this definition of hypothetical chance agreement.

4. Tau (Ma and Redmond, 1995),

$$\tau = \frac{P_c - \sum_{k=1}^{q} \beta_k p_{+k}}{1 - \sum_{k=1}^{q} \beta_k p_{+k}}, \tag{4}$$

where β_k is the user specified *a priori* probability of membership in map class k.

5. User's accuracy for cover type i, the conditional probability that an area classified as category i by the map is classified as category i by the reference data,

$$P_{Ui} = p_{ii}/p_{i+}. \tag{5}$$

6. Producer's accuracy for cover type j, the conditional probability that an area classified as category j by the reference data is classified as category j by the map,

$$P_{Aj} = p_{jj}/p_{+j}. \tag{6}$$

7. Conditional kappa for the map classifications in category (row) i,

$$\kappa_i = \frac{p_{ii} - p_{i+}p_{+i}}{p_{i+} - p_{i+}p_{+i}} = \frac{P_{Ui} - p_{+i}}{1 - p_{+i}}. \tag{7}$$

8. Conditional kappa for the reference classifications in category (column) j,

$$\kappa_j = \frac{P_{Aj} - p_{j+}}{1 - p_{j+}}. \tag{8}$$

These parameters use different information contained in the error matrix (Congalton, 1991) and summarize the error matrix at various levels. For example, P_c, κ, κ_e, and τ provide a single summary measure for the entire error matrix, whereas P_{Aj} and κ_j provide a summary by columns of the error matrix, and P_{Ui} and κ_i provide a summary by the rows of the error matrix. Each of these summary measures obscures potentially important detail contained in the error matrix, so that the full error matrix should be reported whenever possible.

SINGLE SUMMARY MEASURES OF THE ERROR MATRIX

Of the single number summary measures, the most basic is P_c. Other accuracy measures are derived from P_c as the starting point. The motivation for κ and τ is to account for agreement between the map and reference classifications that could be attributable to random chance. These parameters start with an observed measure of agreement, P_c, and then adjust P_c by "a hypothetical expected probability of agreement under an appropriate set of baseline constraints" (Landis and Koch, 1977, p. 163). "Adjusting" for chance agreement is better terminology than "correcting" for chance agreement because the latter has the connotation that P_c is somehow an incorrect representation of accuracy. P_c represents a legitimate probability describing one aspect of map data quality. Users may choose to represent accuracy by another parameter, but it is not relevant to claim that P_c is incorrect or that it provides a biased measure of accuracy.

κ and τ measure the extent to which the observed probability of agreement exceeds the probability of agreement expected under the specified hypothetical baseline constraints. As shown by Ma and Redmond (1995), κ uses marginal proportions of the *observed* map, and τ uses marginal proportions specified *prior* to constructing the map. The hypothetical nature of the adjustment for "chance agreement" needs to be emphasized. In reality, areas classified correctly "by random chance" are indistinguishable from areas classified correctly because of some more favorable aspect of the classification procedure. That is, it is impossible to go to the map and identify those areas (e.g., pixels or polygons) that have been correctly classified by random chance. If the objective is to describe the accuracy of a final map product, the user of this particular map is probably not concerned with the hypothetical proportion of area classified correctly by random chance. Such areas, even if they could be identified, are still classified correctly, and attributing a hypothetical reason for the classification being correct is irrelevant to applications requiring this map. If the overall map accuracy is 80% (P_c=0.80), the user holds a map for which a randomly selected area has an 80% chance of being correctly classified. Thus estimating κ or τ for a final map product is not an informative accuracy mea-

sure, and P_c, P_{Aj}, and P_{Ui} are more relevant accuracy parameters because of their direct interpretation as probabilities characterizing data quality of this particular map.

Reporting κ or τ to summarize a final map product provides a misleading representation of the probability that an area on the map is correctly classified because both κ and τ are always smaller than P_c. This is seen by noting that both κ and τ may be written as $[P_c-a]/(1-a)$, where a is the hypothetical adjustment for chance agreement. Then it follows that $[P_c-a]/(1-a)-P_c=[P_c-a-P_c(1-a)]/(1-a)=a(P_c-1)/(1-a)\leq 0$ because $(P_c-1)\leq 0$ and $(1-a)\geq 0$. Therefore, both κ and τ underrepresent the true probability of a correct classification. A similar argument applies to the relationship between conditional kappa and user's and producer's accuracy; that is, $\kappa_i \leq P_{Ui}$, and $\kappa_j \leq P_{Aj}$.

When adjusting P_c for chance agreement, κ, κ_e, and τ incorporate different adjustments. Which adjustment is best is difficult to discern and will probably remain a matter of debate. In the detailed discussion of Brennan and Prediger (1981, p. 690) of agreement measures, a key distinction they proposed is whether the marginal proportions are considered fixed or free to vary: "A margin is 'fixed' whenever the marginal proportions are known to the assigner before classifying the objects into categories," and "a margin is 'free' whenever the marginal proportions are not known to the assigner beforehand." The identification of the marginal proportions of an error matrix as fixed or free determines the measure of chance agreement used to adjust P_c.

The adjustment for chance agreement used in κ is $\sum_{k=1}^{q} p_{k+}p_{+k}$. Agresti (1996, p. 246) cites the dependence of the κ definition of chance agreement on the marginal proportions as a primary source of controversy on the utility of this measure. The κ adjustment is tantamount to assuming fixed map marginal proportions. But this assumption imposes a circularity in reasoning because the map (row) marginal proportions are the *result* of the classification, not a fixed set of marginal proportions the map construction process was required to match. The adjustment incorporated into κ would be appropriate if the *a priori* row marginal proportions were specified, and the classification were forced to result in a map with those exact marginal proportions. An example from another application illustrates the point. Suppose a physician will evaluate 100 patients, and each patient must be classified into one of five disease categories. The true disease category is known (although not by the physician) for each patient, and the physician is provided with the proportion of patients in each category and told to match those same marginal proportions in his or her evaluation. In this scenario, the κ measure of chance agreement is justified because the marginal proportions provided by the classification are fixed, and the classification (the physician's evaluation) is constrained to match those marginal proportions. The random chance adjustment takes into account the imposed constraint.

Foody (1992) argued that the measure of chance agreement incorporated into κ is not the proper representation for most accuracy assessment problems because the map margins are not fixed, but free to vary. That is, the classification is not constrained to match specified row marginal proportions. In the absence of any information about the land-cover class of a given area, that area would be classified into one of the q classes with equal probability, so $p_{k+}=1/q$. Then assuming independence of the map and reference classifications, chance agreement is still $\sum_{k=1}^{q} p_{+k}p_{k+}$, but substituting $p_{k+}=1/q$ into the equation leads to $(1/q)\sum_{k=1}^{q} p_{+k}=1/q$. This result is the same regardless of how the reference (column) marginal proportions are viewed, either fixed or free, and this chance agreement adjustment results in κ_e. Chance agreement defined for κ_e is smaller than that defined for κ, and this is the basis of Foody's (1992) statement that chance agreement is overestimated by κ.

Ma and Redmond (1995) show that κ_e is a special case of τ, and claim that κ_e is an appropriate measure if unsupervised classification is employed, or if a supervised classification is employed with no *a priori* specification of class membership. In both cases, the row marginal proportions of the error matrix used in the chance agreement adjustment are $1/q$; that is, in the absence of any information about the true class of a particular area of the map, the area will be classified into one of the q land-cover categories with equal probability. If a supervised classification is employed and the *a priori* class membership probabilities specified are not equal, then the measure of chance agreement used by τ is $\sum_{k=1}^{q} \beta_k p_{+k}$, where β_k is the *a priori* probability of classifying an area into category k. In this case, the measure of chance agreement used in τ is based on the premise that some information about the true class of an area exists. That information is contained in the β_k values specified. So now random agreement does not imply the complete absence of prior information about the true class of an area, which is how chance agreement is defined for κ_e.

Although *a priori* probabilities are specified in a supervised classification, the classification procedure is not constrained to match these specified probabilities, so that the measure of chance agreement used in τ is independent of the row marginal proportions of the population error matrix obtained. Ma and Redmond (1995) claim this independence is a desirable feature of τ. But this independence leads to the conceptually disconcerting consequence that two maps having the exact same population error matrices may result in different τ coefficients simply because the *a priori* probabilities are different for the two maps. Further, if the map marginal proportions

are not forced to match β_k, it is unclear how to interpret τ in the Brennan and Prediger (1981) framework. The map (row) marginal proportions are still free to vary, so that perhaps the interpretation is that *if* the map marginals (p_{k+}) had been forced to match the *a priori* probabilities (β_k), then random agreement would be as measured by τ.

Because chance agreement is a hypothetical construct, the various definitions invoked lead to the different accuracy parameters κ, κ_e, and τ. Each parameter assumes different *a priori* information about the true land-cover class of an area, so that it is difficult to claim that one measure is better than another. These measures are simply different. Choosing among these accuracy parameters raises the question of how to represent accuracy, whether to adjust for hypothetical chance agreement at all, and if an adjustment is incorporated, which measure to use. In some cases, all parameters lead to similar conclusions, but in other applications, the conclusions will differ (see first subsection or section on Testing Overall Map Accuracy). Numerous other accuracy measures could be defined, and some are reviewed by Kalkhan et al. (1995; 1996). Bishop et al. (1975) and Agresti (1990) distinguish between measures of *association* and measures of *agreement*, and state that strong association in a contingency table (error matrix) does not imply high agreement. Therefore, measures of association should not be applied to accuracy assessment problems.

FURTHER DISCUSSION OF κ

Because κ has generally been accepted and frequently used to summarize the results of an accuracy assessment, it is worth further exploring the properties of this measure. Dicks and Lo (1990), Fung and LeDrew (1988), Janssen and van der Wel (1994), and Rosenfield and Fitzpatrick-Lins (1986) all state that κ uses all cells of the error matrix, not just the diagonal entries used by P_c. Others apparently disagree. Zhuang et al. (1995, p. 427) stated that κ does "not directly include the effects of off-diagonal entries on the accuracies of individual classification categories and overall classification." Interpreting such conflicting views is difficult, and some of the difficulty is attributable to authors defining terms differently.

The observed marginal proportions of the error matrix are obviously incorporated into κ, so that κ does use some of the off-diagonal information in the error matrix. But different internal configurations of the error matrix p_{ij}'s can result in the same row and column marginal proportions, so that, in that sense, κ does not use all cells of the error matrix. Consider the first two error matrices shown in Table 2. Both have P_c=0.636 and the same marginal proportions. The two error matrices are obviously different internally, but both have κ=0.450. User's and producer's accuracy differ for the two matrices, and the decision of which error matrix represents a better

Table 2. Example Population Error Matrices and Associated Accuracy Parameters

Class	A	*B*	*C*	p_{i+}	P_{Ui}
Error Matrix 1 (P_c=0.636, κ=0.450)					
A	0.2727	0.0000	0.0909	0.3636	0.750
B	0.1818	0.1818	0.0000	0.3636	0.500
C	0.0000	0.0909	0.1818	0.2727	0.667
p_{+j}	0.4545	0.2727	0.2727		
P_{Pj}	0.600	0.667	0.667		
Error Matrix 2 (P_c=0.636, κ=0.450)					
A	0.3636	0.0000	0.0000	0.3636	1.000
B	0.0909	0.1364	0.1364	0.3637	0.375
C	0.0000	0.1364	0.1364	0.2728	0.500
p_{+j}	0.4545	0.2728	0.2728		
P_{Pj}	0.800	0.500	0.500		
Error Matrix 3 (P_c=0.660, κ=0.370)					
A	0.4500	0.1100	0.0400	0.60	0.75
B	0.1500	0.1500	0.0000	0.30	0.50
C	0.0000	0.0400	0.0600	0.10	0.60
p_{+j}	0.60	0.30	0.10		
P_{Pj}	0.75	0.50	0.60		
Error Matrix 4 (P_c=0.660, κ=0.469)					
A	0.3600	0.1000	0.1400	0.60	0.60
B	0.0400	0.2000	0.0600	0.30	0.67
C	0.0000	0.0000	0.1000	0.10	1.000
p_{+j}	0.40	0.30	0.30		
P_{Pj}	0.90	0.67	0.33		
Error Matrix 5 (P_c=0.660, κ=0.490)					
A	0.2644	0.0600	0.0089	0.3333	0.793
B	0.0422	0.1811	0.1100	0.3333	0.543
C	0.0267	0.0922	0.2144	0.3333	0.643
p_{+j}	0.3333	0.3333	0.3333		
P_{Pj}	0.793	0.543	0.643		

classification depends on the relative importance of each land-cover class, and the relative importance of user's and producer's accuracy to the objectives of the particular mapping project. Both P_c and κ obscure class-level differences, and this example illustrates an inherent problem with summarizing the error matrix by a single number.

Another feature of κ is that when the corresponding row and column marginal proportions of the population error matrix are closer to each other, more observed agreement (higher P_c) is needed to attain the same value of κ (Lee and Tu, 1994). That is, if p_{k+} and p_{+k} are similar for each class k, P_c must be higher to achieve the same κ as a map with a greater disparity between p_{k+} and p_{+k}. A highly desirable feature of a land-cover map is for the proportion of area in each land-cover class identified by the map to match the proportion for that class that exists on the ground ($p_{k+}=p_{+k}$). Yet κ penalizes a map achieving this desirable feature. This is demonstrated numerically with the third and fourth error matri-

ces in Table 2. Error matrices 3 and 4 both have P_c=0.66. Error matrix 3 has row proportions (p_{k+}) 0.6, 0.3, and 0.1 and matching column proportions (p_{+k}) 0.6, 0.3, and 0.1, yielding the highly favorable result that area estimates for each land-cover class from the map match the actual areas as given by the reference data. Error matrix 4 has row proportions of 0.6, 0.3, and 0.1 and column proportions of 0.4, 0.3, and 0.3. The classification resulting in error matrix 4 has poorer agreement between the reference and map marginal proportions, yet error matrix 4 has a higher κ (κ=0.47) than error matrix 3 (κ=0.37), even though both maps have P_c=0.66. Error matrix 3 is penalized with higher chance agreement despite possessing a desirable match between row and column marginal proportions.

Consider still another classification represented by error matrix 5 which also has P_c=0.66, and similar to error matrix 3, has matching row and column proportions. In error matrix 5, all map classes are equally represented (1/3), and user's and producer's accuracy are slightly higher than those shown for error matrix 3. With P_c the same as error matrix 3, and user's and producer's accuracy only slightly higher than in error matrix 3, error matrix 5 has a much higher κ (κ=0.49 versus κ=0.37) because it has smaller chance agreement (0.333) than that of error matrix 3 (0.46). Given that error matrices 3 and 5 both possess the desirable feature that the row proportions match the column proportions, it is not clear why chance agreement should differ between the two just because the configuration of marginal proportions is different in the two error matrices.

κ_c resolves some of the confusion in interpreting κ because it is based only on the number of land-cover categories, q, not on the marginal proportions. For κ_c, as q increases, chance agreement ($1/q$) decreases. This is intuitively appealing because if more land-cover categories are added, we would expect fewer correct guesses if the map were classifying area completely at random. When several maps, all with the same number of categories, are compared, κ_c orders these maps from best to worst in exactly the same way as P_c because κ_c is a linear rescaling of P_c. That is, $\kappa_c=a+bP_c$, where $a=1/(1-q)$ and $b=q/(q-1)$, so the difference between κ_c values for the two error matrices is b times the difference between the P_c values. Thus, in this setting, κ_c rescales the magnitude of the difference in accuracy, but does not alter the ordering or ranking.

CLASS-LEVEL ACCURACY MEASURES

When interest focuses on the accuracy for particular land-cover classes, attention shifts to row and column accuracy measures such as user's and producer's accuracy, and conditional kappa. Although the labels user's and producer's accuracy are not universally accepted (cf. Janssen and van der Wel, 1994; Lark, 1995), those terms will be applied to the conditional probabilities defined by Eqs. (5) and (6). User's accuracy is related to commission error, and producer's accuracy is related to omission error (Janssen and van der Wel, 1994).

Rosenfield and Fitzpatrick-Lins (1986) suggested using conditional k for the same reason motivating κ, which is to incorporate an adjustment for hypothetical chance agreement. The relationship between conditional kappa (κ_i) and user's accuracy can be illustrated by writing κ_i as

$$\kappa_i=\frac{p_{ii}-p_{i+}p_{+i}}{p_{i+}-p_{i+}p_{+i}}=\frac{P_{Ui}-p_{+i}}{1-p_{+i}}.$$

Thus conditional κ defines random agreement as p_{+i}, and therefore adjusts user's accuracy by this column proportion for reference class i. Recall that p_{+i} is the true (reference) proportion of area in land-cover class i, not the mapped area proportion. The result of this adjustment is that those land-cover classes that are common in the mapped region must have higher user's accuracy to achieve the same conditional κ as a less common cover class. The justification for this penalty invoked by κ_i to common cover classes seems tenuous. Conditional κ is apparently based on a premise that the map somehow "knows" the proportion of pixels in category i, and therefore the map should assign pixels to that category according to those known proportions. That is, κ_i assumes the margin p_{+i} is fixed in the Brennan and Prediger (1981) framework. If the map indeed has such "knowledge" of p_{+i}, that is a favorable feature of the classification process, and the accuracy of such a map should not be penalized by higher chance agreement. Instead of using p_{+i} to represent random agreement, it seems preferable to use $1/q$ resulting in a conditional κ parameter analogous to κ_c, $(P_{Ui}-1/q)/(1-1/q)$. That is, if the map is truly classifying area completely at random, areas should be assigned to the land-cover classes with equal probability, $1/q$. Similar arguments apply to the relationship between producer's accuracy and κ_j.

AN ILLUSTRATIVE EXAMPLE

A detailed analysis of a published example highlights some of the important issues in defining and interpreting different accuracy measures. This example also demonstrates the confusion that can arise when interpreting different measures. The error matrices presented in Fitzgerald and Lees (1994, their Tables 4 and 5) provide the source material. The two error matrices are based on different classifiers, a neural network (NN) classifier, and a decision tree (DT) classifier. In addition to assessing the overall accuracy of the two classifiers, there is also interest in evaluating and comparing accuracy for the individual land-cover classes.

Based on their analyses, Fitzgerald and Lees (1994, p. 362) expressed a strong preference for κ, and pur-

Table 3. Accuracy Statistics Computed from Fitzgerald and Lees (1994, Table 4) Including the Sea Class

Class	User's Accuracy	Conditional Kappa (row)	Producer's Accuracy	Conditional Kappa (col)	FL Estimates	
					Overall Agreement	$\hat{\kappa}$
1. Dry sclerophyll	0.349	0.346	0.678	0.675	0.992	0.457
2. *E. botroyoides*	0.300	0.299	0.529	0.529	0.998	0.382
3. Lower wet slope	0.026	0.025	0.077	0.075	0.997	0.037
4. Wet *E. maculata*	0.477	0.475	0.372	0.370	0.996	0.416
5. Dry *E. maculata*	0.451	0.449	0.681	0.680	0.997	0.541
6. RF Ecotone	0.593	0.592	0.154	0.153	0.998	0.244
7. Rainforest	0.337	0.336	0.348	0.347	0.998	0.342
8. Paddocks	0.446	0.446	0.962	0.961	0.999	0.609
9. Sea	1.000	0.999	0.992	0.695	0.992	0.820

ported to have demonstrated "that the accepted method of assessing classification accuracy, the overall accuracy percentage [P_c], is misleading especially so when applied at the class comparison level." They further stated that κ is a more "sophisticated measure of interclassifier agreement than the overall accuracy and gives better interclass discrimination than the overall accuracy." Fitzgerald and Lees' (hereafter FL) preference for κ was based on their class-level comparisons. For each land-cover category, they collapsed the full error matrix into a 2×2 table. For example, to estimate their overall agreement proportion for the class dry sclerophyll, they collapsed the entire error matrix into two classes, "dry sclerophyll" and "not dry sclerophyll." The diagonal entries of this collapsed error matrix are then summed and divided by the total sample size to get the estimated overall proportion correct, $\hat{P}_c$. Their $\hat{\kappa}$ statistic is also computed from the collapsed 2×2 table. The FL results are shown in the last two columns of Table 3, and the discrepancy between $\hat{\kappa}$ and $\hat{P}_c$ is apparent.

Estimating P_c from this collapsed error matrix creates one representation of class-level accuracy and is the approach taken by Fleiss (1981, Chap. 13). For the collapsed 2×2 table, P_c represents the accuracy for a dichotomous classification, for example, accuracy of a "dry sclerophyll" and "not dry sclerophyll" classification. The dichotomous classification results in a significant loss of information, and the objective motivating this perspective of class-level accuracy is substantially different from the objective of evaluating the accuracy of the dry sclerophyll class within the context of the nine-category classification represented by the full error matrix. If the objectives call for a two-category classification system, then the FL assessment is appropriate. The class-level $\hat{P}_c$ values computed by FL are high because the sea class dominates the sample size, and this class has extremely high accuracy. This makes any dichotomous classification very accurate. The FL class-level $\hat{\kappa}$ statistics differ greatly from the $\hat{P}_c$ values also because of the dominance of the sea class in the sample. Chance agreement defined by κ will be very high in these collapsed 2×2 tables, so that the $\hat{\kappa}$ values are much smaller than the $\hat{P}_c$ values.

For these same data, user's and producer's accuracy and conditional kappa are computed (first four columns of Table 3). These measures evaluate class-level accuracy within the context of the full nine-category classification scheme. In this representation, a much different conclusion from that presented by FL is obtained. The proportion correct, as measured by user's and producer's accuracy, differs little from the corresponding conditional $\hat{\kappa}$, and the kappa statistic does not provide a better or even different discrimination based on agreement among the classes. The relative rankings of the different classes are exactly the same whether the proportion correctly classified or the kappa statistic is used, and "the disparity in the relative rankings of the overall accuracy values and the Kappa values" noted by FL (p. 366) is an artifact of the collapsed 2×2 tables formed in their definition of class-level accuracy. In general, the FL error matrices do not demonstrate that kappa is "a more rigorous and discerning statistical tool for measuring the classification accuracy of different classifiers" except when class-level accuracy is defined according to their collapsed-class representation. Their conclusions do not generalize to other common representations of class-level accuracy.

Analysis of the error matrices with the dominating sea class excluded (Table 4) provides additional interesting insights into uses of the information in an error matrix. For this eight-category classification (land classes only), the estimated values for P_c are 0.511 for the NN classifier and 0.508 for the DT classifier, and the estimated κ is 0.395 for NN and 0.389 for DT. Although the $\hat{P}_c$ and $\hat{\kappa}$ values differ, both measures suggest little difference between the two classifiers for the overall error matrix. The class-level accuracies are represented by user's accuracy and conditional κ (for rows). The class-level accuracies achieved by the NN and DT classifiers are generally similar, but the accuracy for classes 3 and possibly 8 are sufficiently higher for the NN classifier that this might convey an important practical advantage relative to the DT classifier. Such accuracy differences may be important, but they are not evident from the comparison of the single number summary measures, P_c

Table 4. Accuracy Statistics Computed from Fitzgerald and Lees (1994, Table 4) Excluding the Sea Class for the Neural Network (NN) and Decision Tree (DT) Classifiers

	User's Accuracy		*Conditional Kappa (rows)*	
Class	*NN*	*DT*	*NN*	*DT*
1	0.575	0.613	0.406	0.459
2	0.396	0.383	0.356	0.344
3	0.571	0.407	0.550	0.378
4	0.477	0.432	0.324	0.265
5	0.451	0.420	0.344	0.307
6	0.593	0.554	0.551	0.509
7	0.337	0.355	0.279	0.300
8	0.943	0.879	0.941	0.873

and κ. This illustrates why class-level evaluations are often important.

If the NN classifier is compared to the DT classifier category by category, the ordering of the classifiers is the same whether accuracy is measured by user's accuracy or conditional κ. For a particular classifier, the ordering of the land-cover classes obtained from user's accuracy and conditional κ is not the same. Not surprisingly, differences in the order occur for those classes that have relatively close values of $\hat{P}_{Ui}$ or $\hat{\kappa}_i$.

These examples illustrate that project objectives may dictate that user's accuracy or producer's accuracy for one or more classes is a high priority, and certain misclassifications may be extremely critical while other errors are less important. A more detailed analysis of the error matrix focusing on selected user's and producer's accuracies or particular cell probabilities (p_{ij}) can be done to customize the assessment more closely to project objectives. A weighted kappa statistic (Fleiss, 1981; Naesset, 1996) has been proposed as an overall measure of agreement in which the importance of different misclassifications to the user's objectives can be incorporated into the accuracy measure. This weighting feature is exactly the type of approach that should be employed to link accuracy measures more closely to mapping objectives. Unfortunately, embedding the weighting within the context of a kappa framework results in a measure that suffers from all of the same definition and interpretation problems inherent in κ (see the third section on Further Discussion of κ). Consequently, weighted kappa cannot be recommended as a useful accuracy measure.

TESTING OVERALL MAP ACCURACY

Hypothesis testing may sometimes be required to address accuracy assessment objectives. For example, if a map must have an overall accuracy of P_c=0.70 to meet contractual specifications, a test of the null hypothesis H_0: $P_c \leq 0.70$, can be made against the alternative hypothesis H_a: $P_c > 0.70$. Hypothesis tests can be constructed for P_c, κ, κ_c, τ, or other accuracy parameters. Some general issues pertaining to hypothesis testing in accuracy assessment are reviewed here with the specific focus being tests based on κ.

Testing the null hypothesis H_0: κ=0, evaluates if overall accuracy exceeds that of chance agreement. Fleiss (1981), Agresti (1990), and Janssen and van der Wel (1994) suggest that because map accuracy is anticipated to exceed agreement expected by random chance, testing the hypothesis that κ=0 is often not relevant. A more informative test is to determine if κ exceeds some hypothesized value, say κ_0. The values of κ_0 may be various cutoffs such as those suggested by Landis and Koch (1977) for moderate (0.41–0.6), substantial (0.61–0.8), and almost perfect (0.81–1.0) agreement. For example, a test of the hypothesis that accuracy beyond that expected by random chance may be considered "substantial" translates into testing H_0: $\kappa \leq 0.61$, versus H_a: $\kappa > 0.61$. Fleiss (1981, p. 221) presents the formulas for carrying out such a test.

As with any hypothesis test, the power of the test and practical importance of statistically significant differences found must be considered (Aronoff, 1982). If the sample size is large, the null hypothesis may be rejected even when the observed kappa statistic ($\hat{\kappa}$) does not exceed the hypothesized value (κ_0) by a large amount. For example, Fitzgerald and Lees (1994) reported numerous tests of the null hypothesis H_0: κ=0, and rejected this hypothesis for an observed $\hat{\kappa}$ as small as 0.077. $\hat{\kappa}$=0.077 is clearly not indicative of practically better agreement beyond that expected by random chance, yet this small κ is evidence to claim κ is statistically separable from 0. Because the sample size in the Fitzgerald and Lees (1994) example is large (n=62,727), this particular test is extremely powerful so that even practically unimportant differences from κ=0 will result in rejection of H_0. It should be routine data analysis practice to consider both the statistical significance and practical importance of any hypothesis test result.

Confidence intervals provide important descriptive information and can be used to conduct hypothesis tests. Basic description for accuracy assessment should include estimates of parameters of interest (e.g., P_c, κ, P_{Aj}, P_{Ui}, or κ_i) accompanied by standard errors. An approximate confidence interval is constructed via the formula $\hat{E} \pm z \cdot \mathrm{SE}(\hat{E})$, where $\hat{E}$ is an estimate of the parameter of interest, z is a percentile from the standard normal distribution corresponding to the specified confidence level, and $\mathrm{SE}(\hat{E})$ is the standard error of $\hat{E}$. Both $\hat{E}$ and $\mathrm{SE}(\hat{E})$ depend on the sampling design used to collect the reference data. These confidence intervals assume that the reference sample size is large enough to justify use of a normal approximation for the sampling distribution of $\hat{E}$. For estimates based on the entire error matrix such as $\hat{P}_c$ and $\hat{\kappa}$, sample sizes are likely to be adequately large to satisfy this assumption. For those estimates based on

rows or columns of the error matrix such as $\hat{\kappa}_i$, $\hat{P}_{Ui}$, and $\hat{P}_{Aj}$, sample sizes may be small for rare classes and the normal approximation will not be justified.

Comparison of Error Matrices: Tests Based on a Single Summary Measure

When comparing the accuracy of two or more maps, the primary focus is to rank or order the maps, and to provide some measure of the magnitude of differences in accuracy among the maps. Such comparisons could be based on P_c, τ, κ, or κ_e. Once again, differing opinions have been proffered on which parameter to use. For example, Janssen and van der Wel (1994, p. 424) state that "PCC values [P_c] cannot be compared in a straightforward way" and suggest normalizing the error matrix or using κ to make such comparisons. Although the meaning of "straightforward" is open to interpretation, different error matrices can in fact be compared using P_c, and normalizing an error matrix is shown in the next section to be a questionable analysis strategy. Assuming that the reference samples for the error matrices are independent simple random samples of size n_1 and n_2, so that the estimates $\hat{P}_{c1}$ and $\hat{P}_{c2}$ are independent, a test of H_0: $P_{c1}=P_{c2}$ is obtained from

$$z=\frac{\hat{P}_{c1}-\hat{P}_{c2}}{\sqrt{\hat{P}_{c1}(1-\hat{P}_{c1})/n_1+\hat{P}_{c2}(1-\hat{P}_{c2})/n_2}}, \quad (9)$$

where z is distributed as a standard normal random variable. This is a standard test for comparing two population proportions (cf. Snedecor and Cochran, 1980, Sec. 7.10).

The comparison could also be based on κ (Congalton et al., 1983), and such comparisons have been used in analyses (e.g., Fung and LeDrew, 1988; Gong and Howarth, 1990; Marsh et al., 1994). This test incorporates the adjustment for chance agreement provided by κ. In practice, the same conclusions will often be reached whether P_c or κ is used for the comparison. Congalton et al. (1983) presented estimates for P_c and κ for error matrices derived from four different classification algorithms, and then conducted tests to determine if the κ values for the different population pairs differed (Table 5). The ordering of the four algorithms is different depending on whether P_c or κ is used, but the discrepancy is minor because algorithms 1 and 2 are nearly similar in accuracy. Both accuracy parameters (κ and P_c) reflect this similarity, even though the ordering differs. The z-statistics for the pairwise comparisons of accuracy based on κ and P_c lead to the same general conclusions concerning the statistical significance of differences in accuracy of the different algorithms. Only the comparison between algorithms 1 and 3 might be affected by the choice of parameter depending on the Type I error level chosen.

Jakubauskas et al. (1992) reported P_c and κ values for six different classification methods, all using a four-category land-cover scheme (Table 6). The ordering of the six methods is slightly different depending on whether P_c or κ is used, but the discrepancies again occur when accuracy of two methods is nearly similar to begin with. Rosenfield and Fitzpatrick-Lins (1986, p. 224) reported conditional kappa and user's accuracy for a five-category classification and obtained the same ordering in terms of class-level accuracy from the two measures. The tables presented by Fung and LeDrew (1988) and Dikshit and Roy (1996) provide numerous additional examples for comparing the ordering of maps on the basis of P_c and κ.

When differences in the ordering obtained by P_c and κ among different maps are extreme, this implies potentially important structural differences in the land-cover of the maps being compared. Brennan and Prediger (1981, p. 696) suggest that if the marginals vary from map to map ("assigner to assigner" in their terminology), it is difficult to compare the values of κ that result because accuracy is confounded with chance agreement. That is, comparisons based on κ will yield the same conclusion as comparisons based on P_c unless the marginal proportions (p_{k+} or p_{+k}) of the two error matrices are very different. The question arises, does it make sense to compare numerically maps with such fundamental dif-

Table 5. Accuracy Statistics and Pairwise Comparison Tests for Four Classification Algorithms Reported in Congalton et al. (1983)

Algorithm	n	$\hat{P}_c$	$\hat{\kappa}$
4 modified clustering	632	0.859	0.718
2 nonsupervised 20 clusters	659	0.785	0.586
1 nonsupervised 10 clusters	659	0.766	0.605
3 modified supervised	646	0.714	0.476

z *Statistics for Pairwise Comparisons of Accuracy of the Four Algorithms*

	κ	P_c
1 vs. 2	0.48	0.79
1 vs. 3	3.01	2.17
1 vs. 4	−2.94	−4.32
2 vs. 3	2.43	2.96
2 vs. 4	−3.28	−3.53
3 vs. 4	−5.62	−6.46

Table 6. Accuracy Measures for Six Classification Approaches (Ordered by κ) Reported in Jakubauskas et al. (1992)

Data/Technique	κ	P_c
TM/SPOT supervised	0.618	0.802
TM/SPOT unsupervised	0.606	0.801
SPOT supervised	0.603	0.802
TM unsupervised	0.535	0.785
TM supervised	0.492	0.768
SPOT unsupervised	0.440	0.742

ferences in land cover? For example, if one region has five land-cover classes, all approximately equally distributed, and another region has 10 classes with one class representing 91% of the area, what objective motivates a comparison of accuracy for these very different regions? Would there be reason to expect accuracy of the two regions to be similar? To compare maps with such fundamental structural differences, perhaps some measure of map value defined relative to the user's objectives is a more appropriate basis for the comparison than a measure of map accuracy.

Campbell (1987, p. 351) discusses map accuracy comparisons for the objective of determining which classifications using different images from different dates, classification algorithms, or individuals are best for a given region. In this case, the objective motivating comparison of the maps is clear, but differences in κ are still difficult to interpret. If the same region is being classified, then presumably the same land cover categories are being used, and the reference proportions (p_{+k}) must be the same for each map. But κ also uses the row proportions (p_{k+}) in the adjustment for chance agreement, so that maps constructed from different classification algorithms, interpreters, and dates will have different chance agreement, even though they all classify the same region. The user must decide if it is reasonable to assign these maps different chance agreement, even though they are classifying the exact same region. What is it about the different classification procedures that justifies regarding them as having a different probability of correctly classifying areas by "random chance"? Similar issues were discussed in the second section relative to the definition of chance agreement employed in κ.

κ_c circumvents the confounding problem present in κ attributable to differing marginal proportions. Because chance agreement does not depend on the realized map proportions (p_{k+}), maps classifying the same region using the same land-cover classes, as in Campbell's (1987) application, will have the same chance agreement. If the number of categories is the same for the two error matrices, a test based on κ_c turns out to be equivalent to a test based on P_c. To see this, we begin with the variance for an estimated τ coefficient (the generalization of κ_c) and the z-statistic specified by Ma and Redmond (1995),

$$z=\frac{\hat{\kappa}_{c1}-\hat{\kappa}_{c2}}{\sqrt{\hat{\sigma}_1^2+\hat{\sigma}_2^2}},$$

where $\hat{\kappa}_{c1}$, $\hat{\kappa}_{c2}$ are the estimated κ_c for the two error matrices and $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are the estimated variances for the two sample estimates of κ_c each based on a sample of size n,

$$\hat{\sigma}^2=\frac{\hat{P}_c(1-\hat{P}_c)}{n(1-1/q)^2}.$$

Then substituting the estimates for κ_c and σ^2 into z,

$$z=\frac{\dfrac{(\hat{P}_{c1}-1/q)}{(1-1/q)}-\dfrac{(\hat{P}_{c2}-1/q)}{(1-1/q)}}{\sqrt{\dfrac{\hat{P}_{c1}(1-\hat{P}_{c1})}{n_1(1-1/q)^2}+\dfrac{\hat{P}_{c2}(1-\hat{P}_{c2})}{n_2(1-1/q)^2}}}$$

$$=\frac{(\hat{P}_{c1}-1/q)-(\hat{P}_{c2}-1/q)}{\sqrt{\dfrac{\hat{P}_{c1}(1-\hat{P}_{c1})}{n_1}+\dfrac{\hat{P}_{c2}(1-\hat{P}_{c2})}{n_2}}}$$

$$=\frac{\hat{P}_{c1}-\hat{P}_{c2}}{\sqrt{\dfrac{\hat{P}_{c1}(1-\hat{P}_{c1})}{n_1}+\dfrac{\hat{P}_{c2}(1-\hat{P}_{c2})}{n_2}}},$$

which is exactly the test statistic based on P_c [Eq. (9)].

If two maps have different numbers of classes, the comparison based on κ_c is confounded with a basic structural change in the classification scheme. The user must factor in how these differences in the classification schemes impact the objectives of the project. The decision of which map is better depends on the value of each map to the user, and a map's value will depend both on the map's accuracy, and the relevance of the land-cover classification scheme to the mapping objectives. How P_c and κ_c incorporate this latter component of map value is not clear.

Comparisons based on τ suffer from the same confounding as comparisons based on κ. When using τ, hypothetical chance agreement will differ if the *a priori* probabilities used in a supervised classification differ. In such cases, the comparison is of two fundamentally different classification procedures possessing different *a priori* information. Recall that the measure of chance agreement employed by τ uses hypothetical marginal proportions (β_k) that are not reflected in the actual map proportions. That is, τ adjusts for chance agreement as *if* the map had been forced to match the specified *a priori* proportions of the supervised classification. The user must decide if basing the comparison on a measure adjusting for this *a priori* information is relevant to the objectives.

Finally, an additional complication with comparisons of two maps based on the same reference data should be noted. Even though the two maps may be constructed independently, the reference sample used for the comparison is often the same for both maps, so that the test statistic employed for the hypothesis test should take this lack of independence into account. That is, the test statistics usually provided for comparing κ, τ, or P_c assume independent reference samples, not independent maps, so the independence assumption underlying the statistical comparison is routinely violated. Some type of paired comparison is appropriate if only a single reference sample is obtained, and a test statistic based on an assumption of two independent samples represents at best an approximation to the correct statistical test.

Table 7. Effect of Standardizing an Error Matrix on Various Accuracy Parameters

a) Original Population Error Matrix

		Reference				
		A	*B*	*C*	p_{i+}	P_{Ui}
Map	A	0.271	0.053	0.267	0.591	0.459
	B	0.076	0.027	0.004	0.107	0.252
	C	0.173	0.098	0.031	0.302	0.103
	p_{+j}	0.520	0.178	0.302	$N=225$	
	P_{Aj}	0.521	0.152	0.103		

b) Standardized Population Error Matrix [Values in Bishop et al. (1975) Divided by 100]

		Reference				
		A	*B*	*C*	*Total*	P_{Ui}
Map	A	0.201	0.102	0.697	1.000	0.201
	B	0.474	0.428	0.098	1.000	0.428
	C	0.325	0.470	0.205	1.000	0.205
	Total	1.000	1.000	1.000	$N=225$	
	P_{Aj}	0.201	0.428	0.205		

Standardized Error Matrices

Standardizing or normalizing an error matrix (Congalton, 1991) has been proposed as a way to compare individual cell probabilities of an error matrix because it eliminates the influence of different marginal proportions. Another advantage of standardization cited is that it uses all the information in the error matrix to estimate the cell probabilities. Zhuang et al. (1995) advocate routine use of standardized error matrices.

An example illustrating the result of standardizing an error matrix is presented in Table 7 [data from Bishop et al. (1975), p. 99]. Suppose the original error matrix in Table 7 represents a census of reference data for all $N=225$ pixels in a small region. Summary measures calculated from this error matrix are parameters of the population represented by the map. For example, $P_c=\sum_{k=1}^{q} p_{kk}=0.329$ is the overall proportion of pixels correctly classified in the population. When the population error matrix is standardized, the accuracy parameters differ markedly from those computed from the original matrix. For example, P_c of the standardized error matrix is 0.278 (sum of the diagonal elements divided by 3).

In reality, standardization is applied to a sample error matrix, but evaluating the result of standardizing a population error matrix is relevant for the following reason. An important statistical property of a sample-based estimator of a population parameter is consistency. Cochran (1977, p. 21) defines a method of estimation as consistent "if the estimate becomes exactly equal to the population value when $n=N$, that is, when the sample consists of the whole population." The example calculations for the Table 7 standardized error matrix demonstrate that estimates obtained after standardization are *not* consistent for the parameters of the actual population error matrix. This raises the question of what the parameters estimated following standardization actually represent, and whether these parameters are meaningful to accuracy assessment objectives.

Zhuang et al. (1995) claimed that because user's and producer's accuracies differ, neither is the appropriate estimator of class-level accuracy. If the usual calculations for user's and producer's accuracies are applied to a standardized error matrix, the two accuracies are equal and represented by the diagonal element for that category. This is the effect of standardizing to homogeneous margins. But there is no reason why user's and producer's accuracies should be the same for a particular land-cover class. Story and Congalton (1986) argue that both user's and producer's accuracies may be needed to address project objectives. Both measures represent well-defined conditional probabilities, and this is a compelling reason for retaining them as appropriate accuracy measures. These conditional probabilities are not constrained to be equal, so evaluating both row and column conditional probabilities is part of a thorough analysis of the error matrix. The diagonal probabilities of a standardized matrix must in some sense combine user's and producer's accuracy. But based on the consistency argument, the diagonal cell probability from a standardized error matrix is not a consistent estimator of the parameter p_{kk} of the actual population represented by the map.

Standardization has been employed in contingency table analyses to enhance the interpretability of interaction patterns (see Bishop et al., 1975, examples 3.6-2 and 3.6-3), but the value of standardization to enhance the interpretability or comparability of error matrices in an accuracy assessment setting is questionable. The lack of consistency of estimates from a standardized error matrix is a critical problem. Further, Bishop et al. (1975, p. 97) state that standardization scales the contingency table to fit *hypothetical margins*, which for an error matrix means scaling to hypothetical homogeneous margins. Scaling the error matrix to homogeneous margins is a valid statistical procedure, but the real populations that are the subject of accuracy assessment projects do not have these hypothetical equal margins. Consequently, standardizing leads to estimates of parameters for a hypothetical population that has little relevance to the reality of the accuracy assessment. In their discussions of measures of agreement, neither Agresti (1990), Bishop et al. (1975), nor Fleiss (1981) suggest standardizing a contingency table prior to computing the agreement measures. Unless the parameters estimated from a standardized error matrix can be identified and shown to be relevant to the objectives of accuracy assessment, this procedure should not be used.

SUMMARY

The variety of accuracy parameters available creates a seemingly bewildering array of options from which to choose. Different accuracy measures use different information contained in the error matrix. Selecting an appropriate accuracy measure depends on the objectives of the assessment, which are in turn determined by the objectives of the mapping project. If the objective is to describe the accuracy of a final map product, the overall proportion correct (P_c), user's accuracy (P_{Ui}), and producer's accuracy (P_{Aj}) have a direct probabilistic interpretation in terms of the actual population represented by that map. The appeal of these measures is that they correspond to probabilities of the map user "drawing a correct conclusion from the map (or making a particular type of error) when using it to make a particular prediction" (Lark, 1995, p. 1465). Adjustments for hypothetical chance agreement are unnecessary when the objective is to report the accuracy of a single, final map product, and standardizing an error matrix does not lead to interpretable parameters for the actual population represented by this map.

When the assessment objectives require comparing error matrices, then the choice of an appropriate parameter becomes less clear. If a single summary measure of the error matrix is employed, any of the parameters, P_c, κ, κ_c, or τ is potentially applicable, but none of these parameters directly takes into account specific objectives of a mapping project. Any implication that one accuracy measure is best for all applications is misleading. P_c is the simplest measure to interpret, but if a user wishes to make the comparison adjusting for hypothetical chance agreement, then the question of how to define chance agreement arises. κ, κ_c, and τ each measure chance agreement in different ways, so that a comparison of two maps is obviously dependent on how chance agreement is defined. κ_c results in exactly the same test statistic as P_c if the two maps being compared have the same number of categories. Conclusions from tests based on P_c, κ, κ_c, and τ may differ if one or more of the following occur: 1) The number of land-cover categories in the two maps differs; 2) the land-cover categories themselves differ in the two maps; 3) the marginal proportions (p_{k+} or p_{+k}) differ in the two error matrices (in the case of κ); and 4) *a priori* probabilities (β_k) for each category differ in the two maps in a supervised classification (in the case of τ). In the first three situations, the comparison is of two fundamentally different classification schemes, and the user must ask the question whether it makes sense to employ a statistical comparison to evaluate what is already clearly a different classification scenario. Is a numerical comparison necessary when the two map products being compared represent regions with fundamentally different land cover? This question is relevant regardless of the accuracy measure chosen to make the comparison. The fourth situation provides a comparison based on chance agreement as defined by a feature of the map construction process. Does the user want to base the comparison on a feature of the classification process itself, or on the *outcome* of the classification process? These are the types of questions that should arise when choosing a map accuracy measure for comparing error matrices.

Using a single accuracy parameter to summarize an error matrix may not satisfy the objectives of an accuracy assessment. Aronoff (1982) stated that a single measure of map quality does not provide the information needed to understand the relative advantages of two land-cover maps, and that the error matrix is a valuable tool for such comparisons. Story and Congalton (1986) similarly argued that using only a single value can be extremely misleading, and recommended reporting both user's and producer's accuracies as well as the error matrix. Because it is difficult to anticipate objectives and accuracy needs of all eventual users of a map product, the best course of action is to report the full error matrix along with the sampling design used to collect the reference data. This generally provides sufficient information for each user to estimate and compare accuracy parameters of interest to satisfy the assessment objectives of that project.

This research has been supported by cooperative agreement CR821782 between the Environmental Protection Agency and SUNY-ESF. This manuscript has not been subjected to EPA's peer and policy review, and does not necessarily reflect the views of the Agency. David Verbyla and two reviewers provided several helpful suggestions for improving the manuscript. The Department of Statistics at Oregon State University is acknowledged for supporting this work.

REFERENCES

Agresti, A. (1990), *Categorical Data Analysis,* Wiley, New York.

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

Aronoff, S. (1982), Classification accuracy: a user approach. *Photogramm. Eng. Remote Sens.* 48:1299–1307.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), *Discrete Multivariate Analysis Theory and Practice,* MIT Press, Cambridge, MA.

Brennan, R. L., and Prediger, D. J. (1981), Coefficient kappa: some uses, misuses, and alternatives. *Ed. Psychol. Measure.* 41:687–699.

Campbell, J. B. (1987), *Introduction to Remote Sensing,* Guilford, New York.

Cochran, W. G. (1977), *Sampling Techniques,* 3rd ed., Wiley, New York.

Congalton, R. G. (1991), A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sens. Environ.* 37:35–46.

Congalton, R. G., Oderwald, R. G., and Mead, R. A. (1983),

Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques. *Photogramm. Eng. Remote Sens.* 49:1671–1678.

Congalton, R. G., Green, K., and Teply, J. (1993), Mapping old growth forests on national forest and park lands in the Pacific Northwest from remotely sensed data. *Photogramm. Eng. Remote Sens.* 59:529–535.

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

Dikshit, O., and Roy, D. P. (1996), An empirical investigation of image resampling effects upon the spectral and textural supervised classification of a high spatial resolution multispectral image. *Photogramm. Eng. Remote Sens.* 62:1085–1092.

Fiorella, M., and Ripple, W. J. (1993), Determining successional stage of temperate coniferous forests with Landsat satellite data. *Photogramm. Eng. Remote Sens.* 59:239–246.

Fitzgerald, R. W., and Lees, B. W. (1994), Assessing the classification accuracy of multiresource remote sensing data. *Remote Sens. Environ.* 47:362–368.

Fleiss, J. L. (1981), *Statistical Methods for Rates and Proportions,* 2nd ed., Wiley, New York.

Foody, G. M. (1992), On the compensation for chance agreement in image classification accuracy assessment. *Photogramm. Eng. Remote Sens.* 58:1459–1460.

Fung, T., and LeDrew, E. (1988), The determination of optimal threshold levels for change detection using various accuracy indices. *Photogramm. Eng. Remote Sens.* 54:1449–1454.

Gong, P., and Howarth, P. J. (1990), An assessment of some factors influencing multispectral land-cover classification. *Photogramm. Eng. Remote Sens.* 56:597–603.

Jakubauskas, M. E., Whistler, J. L., Dillworth, M. E., and Martinko, E. A. (1992), Classifying remotely sensed data for use in an agricultural nonpoint-source pollution model. *J. Soil Water Conservation* 47:179–183.

Janssen, L. L. F., and van der Wel, F. J. M. (1994), Accuracy assessment of satellite derived land-cover data: a review. *Photogramm. Eng. Remote Sens.* 60:419–426.

Kalkhan, M. A., Reich, R. M., and Czaplewski, R. L. (1995), Statistical properties of five accuracy indices in assessing the accuracy of remotely sensed data using simple random sampling. In *Proceedings of the 1995 ACSM/ASPRS Annual Convention, ASPRS Technical Papers,* Vol. 1, pp. 246–257.

Kalkhan, M. A., Reich, R. M., and Czaplewski, R. L. (1996), Statistical properties of measures of association and the kappa statistic for assessing the accuracy of remotely sensed data using double sampling. In *Spatial Accuracy Assessment in Natural Resources and Environmental Sciences* (H. T. Mowrer, R. L. Czaplewski, and R. H. Hamre, Eds.), General Technical Report RM-GTR-277, USDA Forest Service, Fort Collins, CO, pp. 467–476.

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

Lark, R. M. (1995), Components of accuracy of maps with special reference to discriminant analysis on remote sensor data. *Int. J. Remote Sens.* 16:1461–1480.

Lauver, C. L., and Whistler, J. L. (1993), A hierarchical classification of Landsat TM imagery to identify natural grassland areas and rare species habitat. *Photogramm. Eng. Remote Sens.* 59:627–634.

Lawrence, R. L., Means, J. E., and Ripple, W. J. (1996), An automated method for digitizing color thematic maps. *Photogramm. Eng. Remote Sens.* 62:1245–1248.

Lee, J. J., and Tu, Z. N. (1994), A better confidence interval for kappa (κ) on measuring agreement between two raters with binary outcomes. *J. Comput. Graph. Stat.* 3:301–321.

Ma, Z., and Redmond, R. L. (1995), Tau coefficients for accuracy assessment of classification of remote sensing data. *Photogramm. Eng. Remote Sens.* 61:435–439.

Marsh, S. E., Walsh, J. L., and Sobrevila, C. (1994), Evaluation of airborne video data for land-cover classification accuracy assessment in an isolated Brazilian forest. *Remote Sens. Environ.* 48:61–69.

Naesset, E. (1996), Use of the weighted Kappa coefficient in classification error assessment of thematic maps. *Int. J. Geogr. Inf. Syst.* 10:591–604.

Rosenfield, G. H., and Fitzpatrick-Lins, K. (1986), A coefficient of agreement as a measure of thematic classification accuracy. *Photogramm. Eng. Remote Sens.* 52:223–227.

Snedecor, G. W., and Cochran, W. G. (1980), *Statistical Methods,* 7th ed., Iowa State University Press, Ames, IA.

Stenback, J. M., and Congalton, R. G. (1990), Using thematic mapper imagery to examine forest understory. *Photogramm. Eng. Remote Sens.* 56:1285–1290.

Story, M., and Congalton, R. G. (1986), Accuracy assessment: a user's perspective. *Photogramm. Eng. Remote Sens.* 52: 397–399.

Treitz, P. M., Howarth, P. J., Suffling, R. C., and Smith, P. (1992), Application of detailed ground information to vegetation mapping with high spatial resolution digital imagery. *Remote Sens. Environ.* 42:65–82.

Vujakovic, P. (1987), Monitoring extensive 'buffer zones' in Africa: an application of satellite imagery. *Biol. Conservation* 39:195–208.

Zhuang, X., Engel, B. A., Xiong, X., and Johannsen, C. J. (1995), Analysis of classification results of remotely sensed data and evaluation of classification algorithms. *Photogramm. Eng. Remote Sens.* 61:427–433.