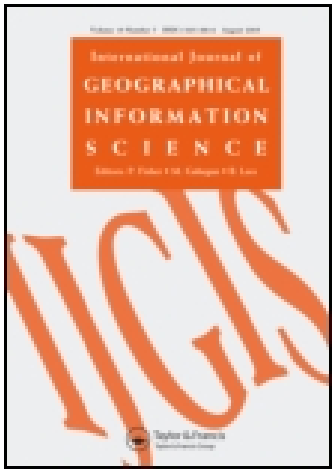

International Journal of Geographical Information Systems
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tgis19

Use of the weighted Kappa coefficient in classification error assessment of thematic maps

Erik Næsset [a]
[a] Department of Forest Sciences , Agricultural University of Norway , P.O. Box 5044, N-1432, Ås, Norway E-mail:
Published online: 24 Oct 2007.

To cite this article: Erik Næsset (1996) Use of the weighted Kappa coefficient in classification error assessment of thematic maps, International Journal of Geographical Information Systems, 10:5, 591-603, DOI: 10.1080/02693799608902099

To link to this article: http://dx.doi.org/10.1080/02693799608902099

INT. J. GEOGRAPHICAL INFORMATION SYSTEMS, 1996, VOL. 10, NO. 5, 591–604

Research Article

Use of the weighted Kappa coefficient in classification error assessment of thematic maps

ERIK NÆSSET
Department of Forest Sciences, Agricultural University of Norway,
P.O. Box 5044, N-1432 Ås, Norway
email: erik.nesset@isf.nlh.no

(*Received 30 June 1994; accepted 12 December 1994*)

Abstract. The weighted Kappa coefficient is applied to the comparison of thematic maps. Weighted Kappa is a useful measure of accuracy when the map classes are ordered, or when the relative seriousness of the different possible errors may vary. The calculation and interpretation of weighted Kappa are demonstrated by two examples from forest surveys. First, the accuracy of thematic site quality maps classified according to an ordinal scale is assessed. Error matrices are derived from map overlays, and two different sets of agreement weights are used for the calculation. Weighted Kappa ranges from 0·34 to 0·55, but it does not differ significantly between two separate areas. Secondly, weighted Kappa is calculated for a tree species cover classified according to a nominal scale. Weights reflecting the economic loss for the forest owner due to erroneous data are used for the computation. The value of weighted Kappa is 0·56.

1. Introduction

Measures of classification accuracy or classification agreement are often used to describe the quality of thematic maps. The results of comparisons between different maps, or between maps and reference surveys, are often presented in square error matrices (contingency tables or agreement matrices). The main diagonal elements of such matrices represent complete agreement. The agreement is frequently expressed by the percentage correct and the Kappa coefficient (Cohen 1960, Congalton and Mead 1983, Congalton *et al.* 1983). The estimator of Kappa is often called Khat (Congalton 1991). Kappa measures the actual agreement minus the agreement expected by chance. Also other metrics that should be used in combination with percentage correct and Kappa have been proposed. A Kappa-like coefficient (Brennan and Prediger 1981) using a somewhat different adjustment of the actual agreement than Kappa, was introduced to the GIS community by Foody (1992). The average mutual information index can be used for assessing the similarity of maps with different themes (Finn 1993).

The Kappa coefficient treats all errors as equally serious. Cohen (1968) has generalized the Kappa measure of agreement to the case where the relative seriousness of each possible disagreement could be quantified (weighted Kappa). Weighted Kappa seems to be largely unknown to the GIS community, although it was briefly mentioned by Rosenfield and Fitzpatrick-Lins (1986).

The approach of weighted Kappa seems useful in two different situations: first, when the categories of thematic maps are ordered. For ordered categories it can

0269–3798/96 $12·00

easily be recognized that the relative seriousness of the different possible disagreement may vary, even though weighted Kappa was introduced in the title of the article by Cohen (1968) as a measure for 'nominal scale agreement'; second, when the categories are classified according to nominal or ordinal scales, and the seriousness of erroneous map data could be quantified for a certain map user or group of map users. The total utility of a map is the sum of the utility obtained for the individual map users. However, the utility of a given map for a specific map user is often unique. Consequently, the seriousness of errors has to be specified individually for specific applications. This can be illustrated by a simple example. Suppose that a thematic map showing agricultural land, coniferous forest, and deciduous forest has been derived from satellite imagery. Misclassification of agricultural land with deciduous forest would not affect the utility of the map for a scientist searching for endangered species living in coniferous forest, while it would be a serious error for a forest owner looking for timber.

It seems plausible to represent the seriousness of misclassifications by the utility loss due to erroneous data. In forest management and management of other natural resources, for example, erroneous map data often induce utility loss due to wrong decisions. The relation between utility loss and accuracy is also addressed in sampling theory (e.g., Cochran 1953 and Hamilton 1978). It is sometimes difficult to quantify the utility of a map for a specific map user. However, for certain applications the utility and the utility loss can be calculated in monetary terms, such as the net present value.

In this paper, the weighted Kappa coefficient is presented. The estimator for the coefficient will be denoted weighted Khat. Formulae for computation of weighted Khat and the estimated variance of weighted Khat based on a sample are given, but the theoretical details are not supplied. These formulae are then applied to two example data analyses. First, error matrices for thematic maps with ordered categories are generated within a GIS environment. Two different approaches to construction of agreement weights are demonstrated. Agreement weights are constructed according to (*i*) a linear scale, and (*ii*) by reflecting the utility loss. Computed weighted Khat is compared to Khat, and the difference between two independent weighted Kappa's is tested. Secondly, an error matrix for data classified according to a nominal scale is analysed using agreement weights reflecting the utility loss. Finally, the effect of various sampling schemes is briefly considered.

2. Methodology

The data should be presented in a square error matrix, which shows how a set of subjects are classified according to the same classification scheme in two different maps.

Assume that n subjects are distributed into k^2 cells by each of them being assigned to one of k categories in one map, and, independently, to one of the same categories in a second map. Let p_{ij} denote the proportion of subjects classified into category i ($i=1, 2, \ldots, k$) in map B and category j ($j=1, 2, \ldots, k$) in map A. The n_{ij}s of the individual cells are supposed to be a sample from a multinomial distribution.

Let

$$p_{i+} = \sum_{j=1}^{k} p_{ij} \tag{1}$$

be the proportion of subjects classified into category i in map B, and

$$p_{+j} = \sum_{i=1}^{k} p_{ij} \tag{2}$$

the proportion of subjects classified into category j in map A. The applied notation is displayed in table 1.

Now, let w_{ij} denote a weight assigned to the i,jth cell, which implies that the proportion p_{ij} in the i,jth cell is to be weighted by w_{ij}. Assume that the weights reflect the agreement. According to a formulation presented by Fleiss *et al.* (1969), the weights are restricted to the interval $0 \leqslant w_{ij} \leqslant 1$ for $i \neq j$, and to be such that the weights representing maximum agreement are equal to 1, i.e., $w_{ii} = 1$. Let

$$p_o = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{ij} \tag{3}$$

be the weighted agreement, and

$$p_c = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{i+} p_{+j} \tag{4}$$

the weighted 'chance agreement'. According to Cohen (1968), weighted Kappa is estimated by

$$\hat{K}_w = \frac{p_o - p_c}{1 - p_c}. \tag{5}$$

It may often be difficult to interpret the strength of agreement associated with Kappa. For most purposes values larger than 0·8 represent almost perfect agreement, values below 0·4 represent poor agreement, and values between 0·4 and 0·8 represent moderate to substantial agreement (Landis and Koch 1977). Weighted Kappa is often multiplied by 100 to give a percentage measure of classification accuracy.

An estimator for the large sample variance of weighted Khat was derived by Fleiss *et al.* (1969). For the ith category of map B, define

$$\bar{w}_{i+} = \sum_{j=1}^{k} w_{ij} p_{+j}, \tag{6}$$

and for the jth category of map A, define

$$\bar{w}_{+j} = \sum_{i=1}^{k} w_{ij} p_{i+}. \tag{7}$$

Table 1. Proportion of subjects distributed into k categories in two maps.

Observed category (map B)	Reference category (map A)					
	1	2	.....	$k-1$	k	Total
1	p_{11}	p_{12}	.....	$p_{1\,k-1}$	p_{1k}	p_{1+}
2	p_{21}	p_{22}	.....	$p_{2\,k-1}$	p_{2k}	p_{2+}
⋮	⋮	⋮		⋮	⋮	⋮
$k-1$	$p_{k-1\,1}$	$p_{k-1\,2}$	.....	$p_{k-1\,k-1}$	$p_{k-1\,k}$	$p_{k-1\,+}$
k	p_{k1}	p_{k2}	.....	$p_{k\,k-1}$	$p_{k\,k}$	p_{k+}
Total	p_{+1}	p_{+2}	.....	p_{+k-1}	p_{+k}	1

The variance may now be estimated by

$$\hat{var}(\hat{K}_w) = \frac{1}{n(1-p_c)^4} \times \left\{ \sum_{i=1}^{k} \sum_{j=1}^{k} p_{ij}[w_{ij}(1-p_c) - (\bar{w}_{i+} + \bar{w}_{+j})(1-p_\sigma)]^2 - (p_\sigma p_c - 2p_c + p_\sigma)^2 \right\}. \quad (8)$$

A formula for the estimated variance for testing the hypothesis H_0: $K_w = 0$ under the assumption of independence, is given by Fleiss *et al.* (1969). It is, however, rarely plausible that agreement is no better than expected by chance. It is usually more important to estimate strength of agreement by constructing a confidence interval for K_w, rather than testing H_0 (Agresti 1990).

It can sometimes be difficult to select the values for the agreement weights. If a particular application of the map is identified and the map data represent a measurable utility for the map user, it is plausible, however, that the weights reflect the utility loss due to misclassifications. Let $U_{c,j}$ be the utility of a subject correctly classified into category j, and let $U_{E,ij}$ be the utility of a subject belonging to category j, which is erroneously classified into category i. Agreement weights reflecting utility loss can be computed by

$$w_{ij} = \frac{U_{E,ij}}{U_{c,j}}. \quad (9)$$

For correct classifications, i.e., $i = j$, $U_{E,ij}$ is equal to $U_{c,j}$. Thus, the weights representing maximum agreement are equal to 1.

If it is impossible to quantify the utility in any measurable units, the selection of values for the weights will probably be rather subjective. Nevertheless, in certain cases the selection of values could be based on previous experiences and rational arguments on the specific topic, analogous to the well known technique of assigning weights to the map layers in suitability analysis based on map overlays.

If the data are classified according to an ordinal scale, and there is no argument claiming that any of the possible disagreements should be treated with special care, it seems reasonable to assign the weights according to a linear scale. Complete agreement is then given the value 1, and the most serious disagreement is given the value 0. Values in the range between 0 and 1 are assigned to the weights of the other cases of disagreement according to a linear scale. This way of assigning weights was applied by Cicchetti and Allison (1971), and may be expressed by

$$w_{ij} = 1 - \frac{|i-j|}{k-1}. \quad (10)$$

Tests on the significance of the difference between two independent weighted Kappas, i.e., the difference between the agreement of two different error matrices, may be performed. This is the same test as the one demonstrated on Kappa by Congalton and Mead (1983). Let $\hat{K}_{w1}$ and $\hat{K}_{w2}$ denote computed weighted Khat for contingency table 1 and table 2, respectively. Let also $\hat{var}(\hat{K}_{w1})$ and $\hat{var}(\hat{K}_{w2})$ be the corresponding estimates of the variance. The test statistic is expressed by

$$Z = \frac{|\hat{K}_{w1} - \hat{K}_{w2}|}{\sqrt{\hat{var}(\hat{K}_{w1}) + \hat{var}(\hat{K}_{w2})}}. \quad (11)$$

Given the null hypothesis H_0: $(K_{w1}-K_{w2})=0$, and the alternative H_1: $(K_{w1}-K_{w2})\neq 0$, H_0 is rejected if $Z\geqslant z_{\alpha/2}$, where $z_{\alpha/2}$ is the appropriate percentile from a standard normal distribution.

3. Example data analysis

3.1. *Example 1: ordinal classification*

Two forest areas in Southeast Norway were surveyed by one surveyor from the Norwegian Institute of Land Survey (NILS) (figure 1(*b*) and figure 2(*b*)) to produce the site quality layer of the official Economic Map Series at the scale 1 : 5000. The Economic Map Series covers almost all productive forest land in Norway. The forest areas are denoted Area 1 and Area 2. Total area of productive forest land is 375 and 234 ha, respectively.

Homogenous site quality polygons were delineated, and the site quality of the individual polygons was determined by field measurements according to the H_{40} site index curves (Braastad 1977, 1980, Tveite 1977). The site index is defined by the age and height of trees, and the specific values of the H_{40} index relate to the tree height at age equal to 40 years. Although H_{40} is basically continuous, the classification was considered to be ordinal, because each subject was classified into one of several mutually exclusive classes, which were predefined to cover specific ranges of the H_{40}.

To assess the accuracy of the site quality layer of the Economic Map Series, two

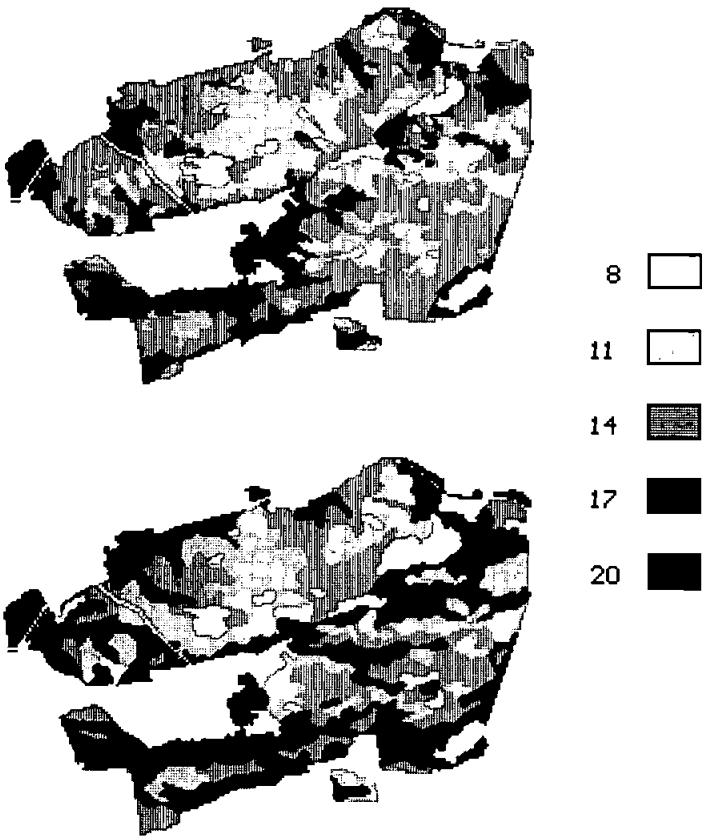

Figure 1. Site quality maps (H_{40} site index, see text) of Area 1. (*a*): Upper, (*b*): lower.

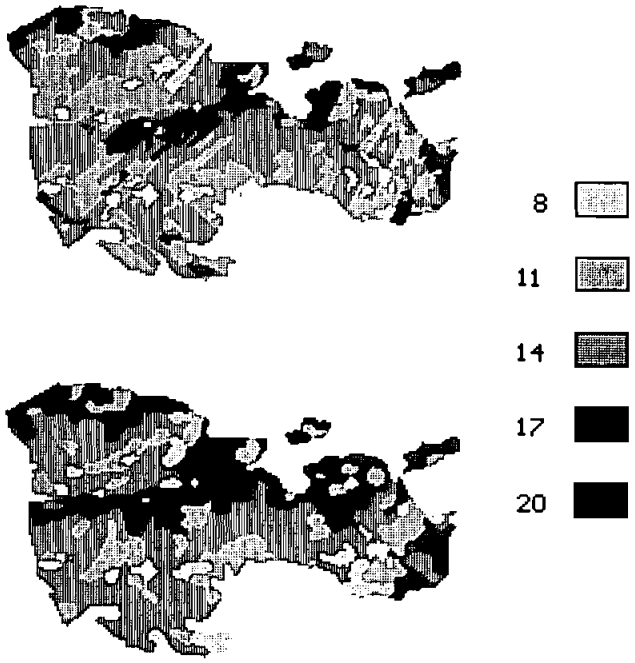

Figure 2. Site quality maps (H_{40} site index, see text) of Area 2. (*a*): Upper, (*b*): lower.

independent reference surveys were carried out by one surveyor from the Glommen Forest Owner Organization (GFOO) (figure 1 (*a*) and figure 2 (*a*)). A higher accuracy for the maps in figures 1 (*a*) and 2 (*a*) than for the maps in figures 1 (*b*) and 2 (*b*) was ensured by dividing the forest areas into smaller and more homogenous polygons.

The site quality classification accuracy was assessed using the grid-based geographical information system IDRISI (Eastman 1992). The four site quality maps based on the field surveys were digitized and stored as separate layers. The site index was assigned to the cell values. The maps of Area 1 are presented in figure 1, and the maps of Area 2 are presented in figure 2.

For each forest area, a layer containing a simple random point sample was generated. By means of overlay analyses and cross-tabulations of Map 1 (*a*) versus Map 1 (*b*) (figure 1) and Map 2 (*a*) versus Map 2 (*b*) (figure 2), the number of points representing areas of correct classification and the number of points representing areas of each category of erroneous classification were derived. These figures were transferred to the SAS package (Sas Institute Inc. 1985), which was used to programme and compute the statistics previously described.

The resulting error matrices are displayed in table 2 and table 3. The small number of observations on the diagonal representing correct classification, clearly indicates large disagreements between the maps.

The classification accuracy was estimated by weighted Khat. The weights were assigned according to a linear scale, which is expressed by (10). The generated agreement weight matrix is shown in table 4. The accuracy was low in both forest areas. The weighted Khat was 0·430 in Area 1 and 0·343 in Area 2 (table 5). Ninety-five per cent confidence intervals for the true values of K_w were computed by $\hat{K}_w \pm 1{\cdot}96\sqrt{\hat{var}(\hat{K}_w)}$.

Table 2. Error matrix for site quality map of Area 1 mapped by GFOO (figure 1 (*a*)) and NILS (figure 1 (*b*)). Number of sample points.

Observed category, site index (figure 1 (*b*))	Reference category, site index (figure 1 (*a*))					
	8	11	14	17	20	Total
8	1	5	3	0	0	9
11	1	55	30	8	0	94
14	0	27	68	8	2	105
17	0	23	74	39	4	140
20	0	0	4	26	26	56
Total	2	110	179	81	32	404

Table 3. Error matrix for site quality map of Area 2 mapped by GFOO (figure 2 (*a*)) and NILS (figure 2 (*b*)). Number of sample points.

Observed category, site index (figure 2 (*b*))	Reference category, site index (figure 2 (*a*))					
	8	11	14	17	20	Total
8	0	8	5	1	0	14
11	2	33	13	2	0	50
14	1	43	51	2	1	98
17	0	6	35	19	8	68
20	0	0	2	3	2	7
Total	3	90	106	27	11	237

Table 4. Agreement weight matrix generated according to a linear scale.

Observed category, site index (Map (*b*))	Reference category, site index (Map (*a*))				
	8	11	14	17	20
8	1·00	0·75	0·50	0·25	0·00
11	0·75	1·00	0·75	0·50	0·25
14	0·50	0·75	1·00	0·75	0·50
17	0·25	0·50	0·75	1·00	0·75
20	0·00	0·25	0·50	0·75	1·00

Table 5. Khat, weighted Khat, their 95 per cent confidence intervals, and test of significant difference between the weighted Kappas computed using agreement weights assigned according to a linear scale.

Area	Khat	Confidence interval	Weighted Khat	Confidence interval	Estimated variance for weighted Khat	*Z* statistic for difference
1	0·282	[0·217, 0·347]	0·430	[0·368, 0·492]	0·00101	1·70 NS
2	0·205	[0·120, 0·289]	0·343	[0·263, 0·422]	0·00163	

Level of significance (H_0: $K_{w1} - K_{w2} = 0$): NS: not significant (>0·05).

The Z statistic in (11) was computed by means of the estimated variances shown in table 5. The null hypothesis was not rejected at the 5 per cent level, which indicates that the accuracy was approximately the same in both forest areas. This may imply that a certain and relatively constant error is present in the site quality layer of the involved map sheets of the Economic Map Series.

The nominal Khat is also supplied in table 5. A correct formula for computation of Khat is provided by Bishop *et al.* (1975) and Hudson and Ramm (1987), among others. The Kappas were significantly smaller than the corresponding weighted Kappas.

To give a second demonstration on how the agreement weight matrix may be used for ordinal classifications, the accuracy was assessed for a forest owner who has to base important forest management decisions on the maps presented. For a forest owner, site quality is an important attribute for computation of the utility of a forest stand, and erroneous classification of a stand induces a utility loss.

Assume that a forest owner's utility function implies maximizing the net present value (NPV) of timber production of a stand. Misclassification of site quality then leads to a reduced NPV due to at least two wrong decisions. First, a sub-optimal treatment schedule is applied to the stand. Secondly, a sub-optimal rotation period is selected. For simplicity, only the effect of the latter was considered. The NPV loss for a stand due to a sub-optimal rotation period can easily be computed given that attributes such as age, trees species, tree heights, number of trees, etc. are available from a forest survey.

The asymmetric agreement weight matrix in table 6 was computed by (9). The net present values were taken from some example stands with different site qualities given by Eid (1990). The rate of interest used for the computation was 4 per cent.

Weighted Khat using the weights reflecting NPV loss in table 6, was 0·553 for Area 1 and 0·472 for Area 2 (table 7). According to the Z statistic computed by means of the estimated variances of table 7, the null hypothesis was not rejected

Table 6. Agreement weight matrix reflecting net present value loss.

Observed category, site index (Map (*b*))	Reference category, site index (Map (*a*))				
	8	11	14	17	20
8	1·00	0·95	0·85	0·70	0·50
11	0·95	1·00	0·95	0·85	0·70
14	0·80	0·95	1·00	0·95	0·85
17	0·60	0·80	0·95	1·00	0·95
20	0·35	0·60	0·80	0·95	1·00

Table 7. Weighted Khat, its 95 per cent confidence interval, estimated variance for weighted Khat, and test of significant difference between the weighted Kappas computed using agreement weights reflecting net present value loss.

Area	Weighted Khat	Confidence interval	Estimated variance for weighted Khat	Z statistic for difference
1	0·553	[0·488, 0·618]	0·00109	1·48 NS
2	0·472	[0·387, 0·557]	0·00189	

Level of significance (H_0: $K_{w1} - K_{w2} = 0$): NS: not significant ($> 0·05$).

($Z = 1{\cdot}48$), which indicates approximately equal accuracy for the two forest areas when economical impacts were included in assessing classification errors.

3.2. *Example 2: nominal classification*

For certain applications of map data the weighted Kappa analysis can be useful for nominal as well as for ordinal classifications, although the approach of assigning weights according to a linear scale is unsuitable for nominal data.

The data used to present a second example were taken from Næsset (1992). Twelve experienced photo-interpreters interpreted tree species in forest stands in a forest area in Southeast Norway. Reference data were collected by intensive field measurements. The forest area was dominated by Norway spruce (*Picea abies* Karst.) and Scots pine (*Pinus sylvestris* L.). A total of 407 observations were presented in an error matrix. The error matrix is displayed in table 8.

The classification accuracy was assessed by assuming that the use of the tree species data was for maximization of the NPV of timber production, as in the previous example. Misclassification of tree species leads to wrong decisions concerning treatment schedules and rotation periods. Only the effect of the latter was considered. For erroneous classification of Norway spruce versus Scots pine, the effect of sub-optimal rotation period on NPV is small, as can be seen from table 9. Table 9 shows agreement weights reflecting NPV loss, and the weights were computed according to empirical example values provided by Eid (1990). The rate of interest was 4 per cent.

Table 8. Error matrix for photo interpretation of tree species.

Observed category (photo-interpretation)	Reference category (field measurement)					
	S	SD	P	PD	M	Total
S	73	21	1	3	16	114
SD	13	32	0	5	39	89
P	0	0	17	48	13	78
PD	1	3	3	28	29	64
M	5	13	2	7	35	62
Total	92	69	23	91	132	407

S = Pure spruce forest; SD = Spruce dominated forest; P = Pure pine forest; PD = Pine dominated forest; M = Mixed conifer forest (spruce and pine)

Table 9. Agreement weight matrix reflecting net present value loss.

Observed category (photo-interpretation)	Reference category (field measurement)				
	S	SD	P	PD	M
S	1·00	0·98	0·92	0·94	0·96
SD	0·98	1·00	0·94	0·96	0·98
P	0·92	0·94	1·00	0·98	0·96
PD	0·94	0·96	0·98	1·00	0·98
M	0·96	0·98	0·96	0·98	1·00

S = Pure spruce forest; SD = Spruce dominated forest; P = Pure pine forest; PD = Pine dominated forest; M = Mixed conifer forest (spruce and pine)

Weighted Khat using the weights in table 9 was 0·558 (table 10), which is significantly higher than the nominal Khat value of 0·322 found by Næsset (1992).

4. Discussion

In both examples weighted Kappa was larger than nominal Kappa. In, for instance, Area 1 of example 1, the nominal Khat value of 0·282 increased to a weighted Khat value of 0·553 using the weight matrix reflecting NPV loss (table 6). A somewhat smaller increase from 0·322 to 0·558 was found in example 2, even though the applied weight matrix of example 2 (table 9) gave larger credit to the misclassifications than the matrix of example 1.

In general, weighted Kappa does not necessarily have to be larger than Kappa, even though weighted Kappa gives some credit to some of the cells representing disagreement. The same weights which generate p_o in (5) also generate p_c. Thus, if the weighted 'chance agreement' is larger than the weighted agreement for cells representing disagreement, weighted Kappa will be smaller than Kappa. The value of Kappa compared to the value of weighted Kappa, therefore depends on the specific data set and the specific weights which are used. Comparison of weighted Kappa from different investigations is meaningful only if the same categories and the same weight matrix are applied.

The selection of values for the weights will in many cases probably be rather subjective. Application of the approach of weighted Kappa seems most relevant when a clear and well defined objective of the accuracy assessment can be stated beforehand. Such a procedure avoids the temptation of manipulating the weights in order to obtain, for example, a high accuracy. One should be aware of this kind of manipulation, particularly if weighted Kappa is used as quality description of commercial map products.

It seems most fruitful to use weighted Kappa in situations where the utility of a map can be quantified in monetary terms for a specific map user, as demonstrated by examples 1 and 2. Various levels of accuracy may be obtained for a particular map, depending on the utility functions of various map users. For some users it is obviously difficult to assess the utility in monetary terms, and the classical problem of economic valuation arises. However, for map applications where the map data represent a well defined economic value, weighted Kappa is more informative as a accuracy measure than nominal Kappa, because weighted Kappa then carries information in the same unit as the decision maker has to consider in the decision making process. This applies particularly to comparison between different thematic maps of the same attribute.

Nominal Kappa applied to data at nominal scales and weighted Kappa applied to data at ordinal scales using weights assigned according to a linear scale, are 'neutral' measures of accuracy analogous to, for instance, difference between means and standard deviations for differences frequently computed for continuous data.

Table 10. Weighted Khat, its 95 per cent confidence interval, and estimated variance for weighted Khat computed using agreement weights reflecting net present value loss.

Weighted Khat	Confidence interval	Estimated variance for weighted Khat
0·558	[0·510, 0·606]	0·00061

Nominal Kappa is, however, just a special case of weighted Kappa, where all the misclassifications are treated equally serious. In cases where it is obvious that the relative seriousness of the different possible errors vary, as demonstrated by examples 1 and 2, one could argue that use of nominal Kappa demands an even stronger justification than the use of weighted Kappa. In spite of the widespread use of Kappa for accuracy assessment of maps and remotely sensed data, justification for treating all the misclassifications as equally serious has hardly been discussed in the literature.

Finally, some general statistical requirements should be considered. The weighted Kappa analysis assumes a multi-nomial sampling model. Only simple random sampling (SRS) fulfils this assumption. Several other sampling schemes than SRS have frequently been used for Kappa analyses of error matrices derived from land cover maps (e.g., Gong and Howarth 1990, Stenback and Congalton 1990, and Agbu and Nizeyimana 1991). The effect of applying other sampling schemes than SRS on weighted Kappa analyses of spatially distributed phenomena has, however, not been reported.

One empirical investigation of some spatial patterns using systematic sampling and stratified systematic unaligned sampling (SSUS), showed that the bias of Khat was negligible (Stehman 1992). The variance of Khat was biased, but the direction of the bias was towards greater precision for both designs compared to SRS. It is difficult to predict the results of a weighted Kappa analysis from the results of a nominal Kappa analysis due to the unequal weights assigned to the individual cells of the error matrix. Nevertheless, it is likely that similar results could have been obtained for weighted Kappa.

Systematic sampling, as well as cluster designs and stratified random sampling, may produce spatially autocorrelated data, which implies that the subjects are not independent. Thus, a basic assumption for t-tests is violated. It has been shown for comparison of means when samples consist of spatially autocorrelated observations, that the t-statistic can be completely misleading (Cliff and Ord 1975). Obviously, the similar Z-statistic in (11) can also be affected by spatial autocorrelation. It should be noted, however, that there are variations of systematic sampling which tend to reduce autocorrelation, such as SSUS.

Also the sample size affects the variance estimates, and thus influences on confidence intervals and the power of tests. The effect of sample size on the confidence interval of the true value of weighted Kappa may be illustrated very simply by assuming that the sample proportions of subjects within all k^2 cells remain unchanged while the sample size increases. From the variance estimator in (8), it can easily be seen that the confidence interval will diminish proportionally by the square root of the number of times the sample size is being increased. Thus, for large samples, even small weighted Kappas will tend to be larger than expected by chance.

5. Conclusions

Weighted Kappa as a measure of nominal as well as ordinal scale classification agreement has been described. When ordinal classification schemes are used, the weighted Kappa analysis applying weights generated according to a linear scale is more informative than the Kappa analysis, because it utilizes the knowledge that the categories are ordered.

The potential use of the weighted Kappa analysis has been demonstrated on error matrices based on real data. The analysis is of particular interest for map applications where the economic loss due to misclassifications can be quantified in

monetary terms. For decision making related to natural resource management, for example, it is an attractive property that the data accuracy and the effect of the decisions may be measured by the same unit. Thus, the weighted Kappa analysis adds a new dimension to the process of thematic map accuracy assessment.

Acknowledgements

The author would like to thank the referees for their helpful and constructive criticisms. This research was funded by the Research Council of Norway, and is a contribution to the research project No. 103478/110.

References

AGBU, P. A., and NIZEYIMANA, E., 1991, Comparison between spectral mapping units derived from SPOT image texture and field soil map units. *Photogrammetric Engineering and Remote Sensing*, **57**, 397–405.

AGRESTI, A., 1990, *Categorical Data Analysis* (New York: John Wiley & Sons, Inc.).

BISHOP, Y. M. M., FIENBERG, S. E., and HOLLAND, P. W., 1975, *Discrete Multivariate Analysis: Theory and Practice* (Cambridge, Massachusetts: The MIT Press).

BRAASTAD, H., 1977, Tilvekstmodellprogram for bjørk. Report 1/77, Norwegian Forest Research Institute, Ås, Norway. (In Norwegian).

BRAASTAD, H., 1980, Growth model computer program for *Pinus sylvestris*. Research Report 35, 265–359, Norwegian Forest Research Institute, Ås, Norway. (In Norwegian with English summary.)

BRENNAN, R. L., and PREDIGER, D. J., 1981, Coefficient kappa: some uses, misuses, and alternatives. *Educational and Psychological Measurement*, **41**, 687–699.

CICCHETTI, D. V., and ALLISON, T., 1971, A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology*, **11**, 101–109.

CLIFF, A. D., and ORD, L. K., 1975, The comparison of means when samples consist of spatially autocorrelated observations. *Environment and Planning A*, **7**, 725–734.

COCHRAN, W. G., 1953, *Sampling Techniques* (New York: John Wiley & Sons, Inc.).

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

COHEN, J., 1968, Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**, 213–220.

CONGALTON, R. G., 1991, A review of assessing the accuracy of classification of remotely sensed data. *Remote Sensing of Environment*, **37**, 35–46.

CONGALTON, R. G., and MEAD, R. A., 1983, A quantitative method to test for consistency and correctness in photointerpretation. *Photogrammetric Engineering and Remote Sensing*, **49**, 69–74.

CONGALTON, R. G., ODERWALD, R. G., and MEAD, R. A., 1983, Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques. *Photogrammetric Engineering and Remote Sensing*, **49**, 1671–1678.

EASTMAN, R., 1992, Idrisi: A Grid based Geographical Analysis Package (Worcester, Mass: Clark University).

EID, T., 1990, Long term forest planning. Economical and biological production possibilities of a forest. Doctor Scientiarum Theses 1990:9, Agricultural University of Norway, Department of Forestry, Ås, Norway.

FINN, J. T., 1993, Use of average mutual information index in evaluating classification error and consistency. *International Journal of Geographical Information Systems*, **7**, 349–366.

FLEISS, J. L., COHEN, J., and EVERITT, B. S., 1969, Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, **72**, 323–327.

FOODY, G. M., 1992, On the compensation for chance agreement in image classification accuracy assessment. *Photogrammetric Engineering and Remote Sensing*, **58**, 1459–1460.

GONG, P., and HOWARTH, P. J., 1990, An assessment of some factors influencing multispectral land-cover classification. *Photogrammetric Engineering and Remote Sensing*, **56**, 597–603.

HAMILTON, D. A., 1978, Specifying precision in natural resource inventories. Rocky Mountain Forest and Range Experiment Station. General Technical Report RM-55, 276–281.

HUDSON, W. D., and RAMM, C. W., 1987, Correct formulation of the kappa coefficient of agreement. *Photogrammetric Engineering and Remote Sensing*, **53,** 421–422.

LANDIS, J. R., and KOCH, G. G., 1977, The measurement of observer agreement for categorical data. *Biometrics*, **33,** 159–174.

NÆSSET, E., 1992, The effect of scale, type of film, and focal length upon interpretation of tree species mixture on aerial photos. *Communications of Skogforsk*, **45,** 1–28. (In Norwegian with English summary).

ROSENFIELD, G. H., and FITZPATRICK-LINS, K., 1986, A coefficient of agreement as a measure of thematic classification accuracy. *Photogrammetric Engineering and Remote Sensing*, **52,** 223–227.

SAS INSTITUTE INC., 1985, *SAS language guide for personal computers.* Version 6 edition, Cary, NC, U.S.A.

STEHMAN, S. V., 1992, Comparison of systematic and random sampling for estimating the accuracy of maps generated from remotely sensed data. *Photogrammetric Engineering and Remote Sensing*, **58,** 1343–1350.

STENBACK, J. M., and CONGALTON, R. G., 1990, Using thematic mapper imagery to examine forest understorey. *Photogrammetric Engineering and Remote Sensing*, **56,** 1285–1290.

TVEITE, B., 1977, Site index curves for Norway spruce (*Picea abies* (L.) Karst.). Research Report 33, 1–84, Norwegian Forest Research Institute, Ås, Norway. (In Norwegian with English summary.)