Explaining the unsuitability of the kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification

Giles M. Foody

School of Geography, University of Nottingham, Nottingham, NG7 2RD, UK

Key words: accuracy, kappa coefficient, chance, prevalence, bias.

Abstract

The kappa coefficient is not an index of accuracy, indeed it is not an index of overall agreement but one of agreement beyond chance. Chance agreement is, however, irrelevant in an accuracy assessment and is anyway inappropriately modelled in the calculation of a kappa coefficient for typical remote sensing applications. The magnitude of a kappa coefficient is also difficult to interpret. Values that span the full range of widely used interpretation scales, indicating a level of agreement that equates to that estimated to arise from chance alone all the way through to almost perfect agreement, can be obtained from classifications that satisfy demanding accuracy targets (e.g. for a classification with overall accuracy of 95% the range of possible values of the kappa coefficient is -0.026 to 0.900). Comparisons of kappa coefficients are particularly challenging if the classes vary in their abundance (i.e. prevalence) as the magnitude of a kappa coefficient reflects not only agreement in labelling but also properties of the populations under study. It is shown that all of the arguments put forward for the use of the kappa coefficient in accuracy assessment are flawed and/or irrelevant as they apply equally to other, sometimes easier to calculate, measures of accuracy. Calls for the kappa coefficient to be abandoned from accuracy assessments should finally be heeded and researchers are encouraged to provide a set of simple measures and associated outputs such as estimates of per-class accuracy and the confusion matrix when assessing and comparing classification accuracy.

1. Introduction

The kappa coefficient of agreement was introduced to the remote sensing community in the early 1980s as an index to express the accuracy of an image classification used to produce a thematic map (Congalton et al., 1983; Rosenfield and Fitzpatrick-Lins, 1986). Early papers highlighted the limitations of conventional approaches to accuracy assessment, especially the omnibus index of overall accuracy that indicates the proportion of correctly classified cases (Turk, 1979). A major concern with the latter is that its magnitude can be highly sensitive to variations in class abundance (i.e. it is prevalence dependent). This problem can be easily illustrated in relation to a basic binary classification such as that used in studies of land cover change. If one class is very rare, as change typically is, an apparently very accurate classification could be achieved by simply allocating all cases to the most abundant class (Fielding and Bell, 1997; Hoehler, 2000). In such circumstances the overall accuracy would seem to be very high but the map produced with the classification would actually provide a very poor representation of the classes, especially with regard to the rare class that may be of particular interest.

To address the problems associated with overall accuracy, the community has been encouraged to estimate and communicate with it measures of per-class accuracy (Story and Congalton, 1986; Janssen and van der Wel, 1994; Congalton and Green, 2009; Stehman and Foody, 2009; Olofsson et al., 2014) as well as explore other measures of accuracy and its reporting (e.g. Finn, 1993; Pontius, 2000; Liu et al., 2007; Foody, 2011; Pontius and Millones, 2011; Comber et al., 2012; Pontius and Parmentier, 2014; Tsutsumida and Comber, 2015; Ye et al., 2018; Ariza-Lopez et al., 2019). For example, the conditional probability that a case has been allocated a class label that corresponds to its actual class of membership which is often referred to as producer's accuracy (Congalton and Green, 2009; Stehman and Foody, 2009; Olofsson et al., 2014) can indicate accuracy on a per-class basis. Similarly, per-class accuracy could be assessed by relating the number of correctly classified cases of a class to the number of cases allocated to that class in the classification and this is often referred to as user's accuracy (Congalton and Green, 2009; Stehman and Foody, 2009; Olofsson et al., 2014). The desire for a single omnibus measure, however, encouraged the exploration of measures of accuracy that seek to summarise accuracy over all classes in a single index and address impacts of issues such as class abundance on the apparent accuracy. Indeed the kappa coefficient was proposed as an index that improved upon overall accuracy (Ubsersax, 1987; Maclure and Willett, 1987) and in the remote sensing community it has been promoted as being an advancement on overall accuracy (Congalton et al., 1983; Fitzgerald and Lees, 1994).

Key arguments put forward for the adoption of the kappa coefficient as an index of classification accuracy were along the lines that it corrected for chance agreement, scales exist for its interpretation, it may be estimated on a per-class as well as on an overall basis and that a variance term may be estimated for it allowing statistically rigorous comparisons to be undertaken (Congalton et al., 1983; Rosenfield and Fitzpatrick-Lins, 1986). Perhaps because of the correction for chance agreement, it is also sometimes claimed that the kappa coefficient is relatively independent of variations in class prevalence (Manel et al., 2001).

The papers that introduced the kappa coefficient for accuracy assessment in remote sensing have had an enormous impact on the research community. These papers have been very highly cited and have been followed by other hugely influential publications that have further promoted the use of the kappa coefficient in accuracy assessment (e.g. Congalton, 1991; Congalton and Green, 2009). These publications have helped to foster the widespread use of the kappa coefficient that has been aided by the inclusion of functionality for its calculation in popular image processing software (Pontius and Millones, 2011).

Despite the widespread promotion of the kappa coefficient and the ease of its estimation, there are many concerns with its use in accuracy assessment. Although widely used, the kappa coefficient has had a troubled history, with concerns ranging from the use of incorrect equations (Fleiss et al., 1969;

Rosenfield and Fitzpatrick-Lins, 1986; Hudson and Ramm, 1987) to more fundamental calls for the kappa coefficient to be abandoned (e.g. Pontius and Millones, 2011). Indeed the use of the kappa coefficient is regarded explicitly as poor practice in accuracy assessment (Olofsson et al. 2013; 2014). Sadly the calls to abandon the use of the kappa coefficient in accuracy assessment seem to have fallen on deaf ears. It may be that the kappa coefficient is still widely used because it has become ingrained in practice and there may be a sense of obligation to use it (Stehman and Foody, 2019). Indeed many researchers seem to use it because precedent for its use exists but given the concerns with the kappa coefficient this is merely an argument to allow mistakes to be repeated. Mistakes happen, but should be used as a positive learning experience that leads to constructive change rather than a situation to be repeated.

It is unclear why the calls to abandon the use of the kappa coefficient in accuracy assessment have not been heeded as the criticisms have been damning with recommendations for good practice clear (e.g. Foody, 1992; Stehman, 1997a; Pontius and Millones, 2011; Stehman and Foody, 2009; Olofsson et al., 2013, 2014). It may be that theoretical arguments have been challenging or that the ease with which the kappa coefficient may be estimated as relevant functionality is often embedded in popular software leads to widespread and possibly unquestioning use. For example, in the period after the publication of the 'death to kappa' paper by Pontius and Millones (2011), the kappa coefficient was reported in half of the relevant literature (Morales-Barquero et al., 2019). The use of the kappa coefficient seems to be embedded into standard practice despite well-known concerns that have been widely disseminated. One possible reason for this unsatisfactory situation is that the community is unaware of the magnitude of the problems associated with the use of the kappa coefficient. Hence, this article aims to revisit major concerns with the use of the kappa coefficient to demonstrate its unsuitability as an index of classification accuracy in remote sensing using simple examples with a focus on highlighting the challenges of interpreting a kappa coefficient by stressing the difficulties in interpreting its magnitude. It will be stressed that all of the arguments put forward for the use of the kappa coefficient are flawed or, in the sense that they are not unusual or unique, irrelevant. The article will first review the estimation of the kappa coefficient and key attributes that have been espoused in support of its use. The latter will be critically evaluated to highlight key concerns before providing some simple examples to demonstrate the problems that can be encountered in the interpretation of the magnitude of a kappa coefficient. Throughout the focus is on commonly encountered situations and hence limited to evaluations of standard hard classifications.

2. Estimation of the kappa coefficient

The kappa coefficient can be estimated easily from the confusion or error matrix that is widely used in classification accuracy assessment. For ease of discussion, the main focus will be on the simplest case of a binary confusion matrix which is widely used in, for example, studies of land cover change (Figure 1). The approach readily extends to larger, multi-class, matrices and this is briefly discussed for completeness. For ease of presentation, it will also be assumed throughout that the sample of cases used to form the confusion matrix was acquired using simple random sampling unless stated otherwise; different sampling designs can be used and the correct formulae for use with them are provided in the literature (e.g. Stehman, 1996, 1997b).

In a binary classification there are just two classes. Thus, in the map produced by a binary image classification, each case (e.g. image pixel) either has (+) a particular trait associated with it or it has not (-). For example, in a remote sensing application the case might be labelled in the map as representing an area of change or of no change. Similarly, the labels might be forest and non-forest or urban and non-urban or to some other specific class of interest or not. Critically, a case may also have similar labels applied to it in a ground reference data set used to assess classification accuracy. The cross-tabulation of the class labels observed in the map and those in the reference data set yields a basic 2 x 2 confusion matrix, often referred to as an error matrix, from which a range of summary measures of classification accuracy can be obtained (Figure 1). Based on the assumption that the map and reference data sources are considered to be two independent raters, the kappa coefficient of agreement may be estimated from this matrix.

Reference ↓

Map →

Class	+	-	$\sum$
+	a	b	$n_{+\cdot}$
-	c	d	$n_{-\cdot}$
$\sum$	$n_{\cdot+}$	$n_{\cdot-}$	n

Figure 1. The confusion matrix for a binary classification based on a simple random sample of *n* cases.

Before exploring the estimation of the kappa coefficient further it may be useful to focus first on the composition of the confusion matrix. The binary confusion matrix has four elements that summarise every possible scenario of class labelling. The number of cases with each of the four possible class allocation scenarios, *a*-*d*, are inserted into the appropriate matrix elements. Of these, *a* cases are labelled as having the trait of interest in both the image classification that forms a thematic map and the reference data; these are often termed true positives. The *d* cases that are labelled as not having the trait of interest in both the image classification and the reference data lie in the other element of the matrix's main diagonal; these are often termed true negatives. Thus, the cases lying in elements of the main diagonal, *a* and *d*, represent those that have been correctly classified. All of the cases that have been incorrectly classified lie in the off-diagonal elements of the matrix. Of these, *b* are those cases that have been classed as having the trait of interest but do not actually possess it; these are commonly referred to as false positives. Such cases represent commission errors, sometimes referred to as type I errors although the use of this terminology can sometimes be problematic (Thron and Miller, 2015). Finally, *c* cases have the trait of interest in the reference data but were classified as not having it; these are commonly referred to as false negatives. These latter cases represent omission errors, sometimes referred to as type II errors. The cases on which the classification and reference data differ in labelling are the misclassifications or errors. In Figure 1, omission is assessed with a focus on the columns of the matrix while commission is assessed with a focus on the rows of the matrix. The total number of cases lying in each row and each column can be determined by summing the relevant matrix elements. These row and column total values are often referred to as the matrix marginal values. Their total, calculated over all rows or all columns, also equates to the total number of cases, *n*, used to form the matrix. The difference between the row and column proportions for a class indicate non-site specific accuracy and indicate map bias which is sometimes referred to as quantity disagreement (Pontius and Millones, 2011; Stehman and Foody, 2019). Finally, the prevalence, θ, of the trait of interest which indicates its abundance may be estimated from $\frac{(a+c)}{n} = \frac{n_{+}}{n}$ and is a property of population being studied. Ideally, a measure of accuracy should reflect only the quality of the classification and not vary with prevalence. Indeed, the prevalence dependency of overall accuracy noted at the beginning of this article is one of its major limitations as a measure of accuracy. Some measures, such as producer's accuracy, are prevalent independent if the diagnostic ability of the classifier is unaffected by prevalence, which can aid their interpretation; in common remote sensing applications the producer's accuracy may, however, be expected to be prevalent dependent.

Using notation similar to Cohen (1960), the kappa coefficient of agreement, κ, is estimated from:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \qquad (1)$$

where p_o is the proportion of cases correctly classified (i.e. overall accuracy) and p_e is the expected proportion of cases correctly classified by chance; note with this notation the distinction between parameters and estimated parameters is not explicit but the text will indicate where sample-based estimates are being made or used. The magnitude of κ lies on a scale from -1 to +1 but interest is typically focused on only on positive values because negative values indicate a level of agreement less than that due to chance and can be difficult to interpret (Sim and Wright, 2005). The maximum value of +1 occurs when there is perfect agreement and a value of 0 arises when the observed agreement equals that due to chance (Cohen, 1960). Commonly the magnitude of the kappa coefficient is interpreted relative to a scale. One such interpretation scale that has been widely used in remote sensing applications is that proposed by Landis and Koch (1977).

Central to the estimation of the kappa coefficient is the estimation of the level of agreement and also the level of agreement that occurs due to chance. For the simple case of a binary confusion matrix such as shown in Figure 1, the proportion of agreement, p_o is estimated from

$$p_o = \frac{a + d}{n} \tag{2}$$

in which a and d are the number of cases correctly labelled (i.e. the true positive and true negative cases), lying in the elements of the main diagonal of the confusion matrix (Cohen, 1960; Congalton et al., 1983). Thus, p_o is simply the sum of all correctly classified cases divided by the total number of cases used to form the matrix and expresses the proportion of correctly labelled cases (i.e. overall accuracy); it is often multiplied by 100 and expressed as a percentage which is commonly termed the percentage correctly classified cases. Although an imperfect index of accuracy, the proportion of correctly allocated cases is relatively easy to estimate and understand (Pontius and Millones, 2011). Before going into any further detail one thing to note at this stage of the discussion is that the kappa coefficient is estimated from p_o, it is an additional analytical step required after the estimation of overall accuracy.

There are a variety of ways to estimate chance agreement (Byrt et al., 1993), but the version that is adopted commonly in remote sensing, which is used in the estimation of Cohen's kappa coefficient, is based on a simple analysis of the row and column marginal values (Byrt et al., 1993; Lantz and Nebenzahl, 1996; Hoehler, 2000, Sim and Wright, 2005). In this, the proportion of agreement expected due to chance, p_e, may obtained from equation 3.

$$p_e = \left(\left(\frac{a+c}{n}\right)\left(\frac{a+b}{n}\right)\right) + \left(\left(\frac{b+d}{n}\right)\left(\frac{c+d}{n}\right)\right) \tag{3}$$

Chance may be modelled differently yielding alternatives to equation 3 and these may be used in equation 1 to yield other indices of agreements. For example, Scott's pi, π, is estimated from equation 1 but, as it is based on different assumptions to the kappa coefficient, the estimation of p_e is different (Byrt et al., 1993; Banerjee et al., 1999).

To illustrate accuracy on a per-class basis it is possible to estimate the conditional kappa coefficient (Rosenfield and Fitzpatrick-Lins, 1986; Czaplewski, 1994; Congalton and Green, 2009). For the class *i*, which has either the + or – label, the latter may be estimated from

$$\kappa_i = \frac{n n_{ii} - n_{i\cdot} n_{\cdot i}}{n n_{i\cdot} - n_{i\cdot} n_{\cdot i}} \qquad (4)$$

The variance for kappa may be estimated (Congalton et al., 1983; Congalton and Green, 2009) and can be usefully expressed in terms of the standard error, σ_κ, which is the square root of the variance. The details of the estimation are not central to the argument in this article but the equation for its estimation for those interested is given in Figure 2. A large literature discusses the estimation of the variance and related terms in more detail (e.g. Fleiss et al., 1969, 2013; Hudson and Ramm, 1987; Czaplewski, 1994).

The standard error may be used to define confidence limits around the estimated value of a kappa coefficient. For example, the 95% confidence interval (95% CI) would be $\kappa \pm 1.96\sigma_\kappa$ as at this level of confidence the standard score, *z*, is 1.96. The statistical significance of a kappa coefficient may also be assessed, using:

$$z = \frac{\kappa}{\sigma_\kappa} \qquad (5)$$

which indicates the degree to which the level of agreement observed is better than that arising from chance alone (Congalton and Green, 2009; Fleiss et al., 2013). More usefully, this also provides the basis to compare an estimated kappa coefficient against other values and also to compare the difference between two estimated kappa coefficients. This is particularly useful when seeking to undertake a statistically rigorous and credible comparison of the accuracy of two thematic maps. For example, two maps, A and B, may have been produced for a region using two different classifiers and the researcher may be interested in knowing if they differ in accuracy. The test for the significance of the difference between two kappa coefficients estimated using independent samples is:

$$z = \frac{\kappa_A - \kappa_B}{\sqrt{\sigma_{\kappa A}^2 + \sigma_{\kappa B}^2}} \tag{6}$$

where κ_A and κ_B are the estimated kappa coefficients for maps A and B respectively, and $\sigma_{\kappa A}$ and $\sigma_{\kappa B}$ are the associated estimates of the standard error of kappa for maps A and B respectively (Cohen, 1960; Congalton and Mead, 1983; Congalton et al., 1983; Rosenfield and Fitzpatrick-Lins, 1986, Smits et al., 1999). Two maps would be deemed to be of different accuracy if $|z|>1.96$ at the 95% level of confidence. If the hypothesis under test has a directional component (e.g. that one map is more accurate than another) a one-sided rather than two-sided test can be undertaken in the usual way (Foody, 2009; Fleiss et al., 2013).

The discussion in this article is focused on binary classifications for ease but the issues extend to multi-class classifications. For multi-class classifications the nature of the confusion matrix and key equations are given in Figure 2.

Reference ↓

	Class	1	2	··	m	$\sum$
Map →	1	p_{11}	p_{12}	··	p_{1m}	$p_{1\cdot}$
	2	p_{21}	p_{22}	··	p_{2m}	$p_{2\cdot}$
	⋮	⋮	⋮	··	⋮	⋮
	m	p_{m1}	p_{m2}	··	p_{mm}	$p_{m\cdot}$
	$\sum$	$p_{\cdot 1}$	$p_{\cdot 2}$	··	$p_{\cdot m}$	1

$$p_o = \sum_{i=1}^{m} p_{ii} \qquad p_e = \sum_{i=1}^{m} p_{i\cdot} p_{\cdot i}$$

$$\sigma_\kappa = \frac{1}{(1-p_e)\sqrt{n}} \sqrt{p_e + p_e^2 - \sum_{i=1}^{m} p_{i\cdot} p_{\cdot i} (p_{i\cdot} + p_{\cdot i})}$$

Figure 2. The confusion matrix for a multi-class classification involving *m* classes, expressed as proportions, together with key equations for the estimation of the kappa coefficient and its standard error.

3. Challenging the arguments for the use of the kappa coefficient

Before addressing the substantive problems with the kappa coefficient it should be noted that a range of problems have been encountered in its use in remote sensing. For example, there is often a failure to recognise impacts of the sample design used to acquire the cases used in estimation (Stehman, 1996), incorrect variance equations have been used (Rosenfield and Fitzpatrick-Lins, 1986), and many comparative assessments have used related rather than independent samples (Foody, 2004) or not recognised the directionality of the study which may require testing for dissimilarities related to inferiority, superiority or equivalence rather than just a difference (Foody, 2009). Similar concerns could be flagged in relation to other indices of accuracy and so such problems are not the central issue of concern to this article. Here, the concern is that the kappa coefficient is unsuitable for use in accuracy assessment, the additional problems encountered in practical application are of very secondary importance. Consequently, the latter are not discussed further especially as such methodological errors are often easy to address with, for example, equations for use with stratified samples (Stehman, 1996) and cluster samples (Stehman, 1997b) as well as statistical tests for related samples (Donner et al., 2000; Foody 2004; 2009; Fleiss et al., 2013).

Central to this article are fundamental problems with the use of the kappa coefficient as an index of classification accuracy. A variety of arguments can be raised against the use of the kappa coefficient in accuracy assessment. These range from the fundamental issue that as a measure of inter-rater agreement it is not a measure of accuracy (Nishii and Tanaka, 1999; Vach, 2005; Wu et al., 2007) to substantial difficulties in its interpretation (Byrt et al., 1993; Lantz and Nebenzahl, 1996; Sim and Wright, 2005; Pontius and Millones, 2011). Here, the central focus is directed at challenging each of the arguments that have been put forward to promote the use of the kappa coefficient in order to highlight its unsuitability as a measure of classification accuracy, summarised in Table 1.

Table 1. A summary of the seven main arguments offered for the adoption of the kappa coefficient and a brief critique of each, highlighting the argument to be either seriously flawed or irrelevant, in the sense that while it may be a valid statement there is nothing unusual or different to other standard, often simpler, indices of accuracy. In short, not a single one of the key arguments put forward for the use of kappa has any real merit, each is either deeply flawed or equally applicable to other indices.

Arguments for the use of kappa	**Reality**
It ‘corrects’ for chance agreement	Flawed argument. There is no need to ‘correct’ for chance agreement. The source of error is unimportant in the assessment of classification or map accuracy. Furthermore, chance is an artificial construct and the way it is modelled in the estimation of κ is inappropriate.
Its estimation is based on the entire confusion matrix	Flawed argument, indeed one that is completely untrue. The estimation is actually based on the main diagonal together with the row and column marginal totals.
It can be estimated on an overall and per-class basis	Irrelevant as the exact same can be argued for other standard measures of accuracy such as overall accuracy (i.e. the proportion of cases correctly classified) with per-class statements from the user’s and producer’s perspectives.
It is, to a large degree, prevalent independent	Flawed argument as untrue. Kappa is, like many other indices, very dependent on class prevalence.
A variance term may be estimated for it.	Irrelevant as the exact same can be argued for other standard measures of accuracy such as the proportion of cases correctly classified.
It allows rigorous comparison of estimates of classification accuracy.	Irrelevant as the exact same approach to comparison, which requires variance estimates, can be used with other measures of accuracy. The commonly promoted approach is also suitable for situations in which independent samples are used but often the same sample is used; methods for the comparison of accuracy estimates obtained from the same sample are available. The comparison of kappa coefficients is also problematic if there are differences in prevalence.
Scales exist for its interpretation	Flawed argument. A variety of scales exist but any scale is arbitrary and cannot be expected to be of universal applicability. The scales also ignore problems linked to issues such as class prevalence.

The kappa coefficient is designed for application to data arising from two independent raters and provides a measure of the degree to which they agree in labelling. Indeed, an early article introducing the kappa coefficient to the remote sensing community focused on its use as a measure of inter-rater agreement (Congalton and Mead, 1983). However, this type of analysis is not the scenario encountered in the assessment of classification accuracy, notably because the ground reference data are supposed to represent the true condition and the desire is to yield a measure of accuracy not simply agreement.

Classification accuracy is a measure of the quality with which a set of cases have been labelled. Fundamentally, the concern in accuracy assessment is with the amount of error or mis-labelling that has occurred in the classification. In this way the accuracy assessment is useful in terms of assessing the fitness for purpose of the classification. The latter would typically require a comparison of the estimated accuracy relative to some target value that indicates the minimum acceptable accuracy for the proposed use of the classification. A target accuracy should ideally be defined before the classification is undertaken and be tailored to the specific purpose of the classification (Foody, 2008). For example, in the pioneering work linked to Anderson (1971) and Anderson et al. (1976) for the mapping of broad land cover classes over a large area, a target of 85% correct allocation with the classes mapped to approximately equal accuracy was used. This target value was well-justified for the specific application and data sets used. For a different mapping application, a target for the specific needs of that individual application should be defined and used; the 85% target put forward by Anderson et al. (1976) is not a universally applicable one. For example, a simple binary classification involves fewer classes than the application Anderson et al. (1976) addressed and a higher target accuracy might be appropriate. An example used below, for instance, sets a target that comprises an overall accuracy of 95% with the producer’s accuracy for the two classes to be at least 95%. Key attractions of this sort of Anderson-type target are that a target value can be defined in advance of the classification and it may, to some extent, help to address concerns with prevalence dependency. The latter arises because the target includes the producer’s accuracy for each class and this measure of accuracy is independent of prevalence if the diagnostic ability of the classifier is fixed (Rogan and Gladen, 1978; Maclure and Willetts, 1987); but note that the valuable attribute of prevalence independence is lost if the ground data set is imperfect (Foody, 2010) or if the diagnostic ability of the classifier changes with prevalence.

The desire for a target highlights an initial problem with the use of the kappa coefficient: how can a sensible target value be defined in advance of a mapping study when the marginal values of the confusion matrix are unknown? In brief, it will typically be infeasible to define a meaningful kappa coefficient as a target value in advance of the classification. It could be argued, however, that a target value is not required with the use of the kappa coefficient as the quality of the classification can be assessed relative to an interpretation scale. This will be one of the problems with the kappa coefficient that will be discussed below.

As highlighted in the introduction, several key attributes have been routinely suggested as arguments for the use of the kappa coefficient in the assessment of classification accuracy. Perhaps the most widely used argument for the adoption of the kappa coefficient is, essentially, that it corrects for chance agreement. Although the exact meaning of ‘chance correction’ is not always clear the core thrust appears to be that it adjusts the assessment for the effect of chance agreement; the kappa coefficient essentially quantifies the level of agreement beyond that due to chance. This is an important observation as the kappa coefficient is often treated as a measure of overall agreement rather than a measure of agreement beyond chance (Jiang and Liu, 2011) and, as noted above, chance may be modelled in different ways and so needs to be quantified with care. Because of the assessment being made relative to a random classification, which is unrealistic of real land cover mosaics, the kappa coefficient fails to meet the map relevant criterion for good practice (Stehman and Foody, 2019). Moreover, the aim of an accuracy assessment is, essentially, the estimation of how much error has occurred; the lower the error the greater the accuracy. Note the origin of the error or the reason for correct labelling is of absolutely no concern to the measurement of accuracy. In a conventional accuracy assessment, a map label is either correct or it is not. There may well be interest in understanding error, especially as a means to further enhance a classification-based analysis, but such assessments of skill require a different type of analysis (Turk, 1979); a distinction between the assessment of classifier performance that indicates diagnostic ability and the assessment of classification accuracy is required (Turk, 2002). Accuracy assessment merely seeks to quantify the amount of error, the origin or source of the error is irrelevant. There is, therefore, no interest in chance agreement and no desire to correct for it in a standard accuracy assessment. Indeed rather than estimate and remove the chance agreements the community should regard such agreements as a windfall gain (Turk, 2002). Even if there was a desire to explore the issue of chance agreement the estimation of its magnitude for the calculation of the kappa coefficient, equation 3, is inappropriate. Since the ground reference data represent reality rather than labels from another independent rater, it may be more appropriate to have fixed column marginal values determined by the number of classes with $p_e = 1/m$ (Brennan and Prediger, 1981; Foody, 1992).

Another popular argument for the use of the kappa coefficient is that its variance may be estimated which facilitates rigorous testing. In particular, the ability to obtain the variance for kappa allows tests of the statistical significance of the difference between two kappa coefficients to be undertaken (Rosenfield and Fitzpatrick-Lins, 1986; Congalton and Green, 2009). These arguments are well-founded and the ability to rigorously compare estimates is a useful attribute. This situation is, however, nothing particularly special to the kappa coefficient. The variance of other estimates of accuracy such as the overall accuracy, which is simply a proportion (p), can also be calculated. The standard error for a proportion, assuming the use of a simple random sample, can be estimated from:

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}} \tag{7}$$

Thus, the variance and related statistics can be obtained for proportions (Fleiss et al., 2013) such as overall, producer's and user's accuracy. Furthermore, contrary to claims to the reverse (Jansen and van der Wel, 1994), it is possible to rigorously compare estimates of the proportion of correctly classified cases. Thus, the statistical significance of the difference in the accuracy of two classifications could be assessed using overall accuracy. The assessment would be similar to that indicated by equation (6) but with the proportion correct, p_o, and its associated variance term, which can be expressed as the standard error, σ_P, for each classification used instead of the kappa coefficients and their standard errors (Stehman, 1997a; Foody, 2004):

$$z = \frac{p_{o_A} - p_{o_B}}{\sqrt{\sigma_{pA}^2 + \sigma_{pB}^2}} \quad (8)$$

Equation 8 allows the statistical significance of differences in proportions, such as overall accuracy, on the assumption that the samples used are independent. Often in remote sensing applications the same ground reference data set is used and the effect this has on the analysis could be addressed by integrating a covariance term into the test or by adopting a test suited for use with related samples such as the McNemar test as an alternative (Foody, 2004; 2009).

The ability to estimate a measure of accuracy on a per-class basis has also been highlighted as an advantageous feature associated with the kappa coefficient. Often referred to as conditional kappa this allows assessment on a class-specific rather than overall basis. Although this is a useful feature it is also nothing special or unique to the kappa coefficient. As noted above, per-class measures of accuracy can be obtained directly from the confusion matrix used to estimate p_o. For example, simple per-class measures such as user’s and producer’s accuracy can be obtained by analysing the relevant row and column of the confusion matrix depending on whether errors of commission or omission are important. For example, the producer’s accuracy (P) for the class with the trait of interest is estimated from $P_+ = a/n_{\cdot+}$; often referred to as the true positive rate, recall or sensitivity. Similarly, the producer’s accuracy may be calculated for the class without the trait of interest from $P_- = d/n_{\cdot-}$; often referred to as specificity. Alternatively, with a focus on commission error, the user’s accuracy (U) may be calculated for each class. For example, the user’s accuracy for the class with the trait of interest may be estimated from $U_+ = a/n_{+\cdot}$; often referred to as the positive predicted value or precision although this latter term should perhaps be avoided due to the potential for mis-interpretation. Sometimes researchers combine measures to yield a single summary indicator of classification accuracy. One such measure which utilizes the producer's accuracy for each class is Youden's *J* which is estimated as $J = P_+ + P_- - 1$ (Allouche et al., 2006; Hand, 2012); sometimes referred to as the true skills statistic or informedness. This latter index is sometimes attractive as an overall summary measure of classification accuracy as the components may be prevalent independent if the diagnostic ability of the classifier is fixed and, although not without concerns, its variance may also be estimated (Allouche et al. 2006). However, there are many measures of accuracy and these can be combined in various ways. For example, average accuracy or the F1 score can be estimated. Such measures, however, are challenging to interpret and of questionable value (Stehman and Foody, 2009; Liu et al., 2007). Indeed many measures of accuracy are available and may be sensitive to different things (Hand, 2012). For a statement of map accuracy to be useful the error measure adopted should be justified and appropriate to the task in-hand (Fielding and Bell, 1997).

A key feature often used in the promotion of the use of the kappa coefficient in accuracy assessment is that scales to interpret the kappa coefficient are available. The existence of a meaningful scale could also be argued to remove the common desire for a target value in accuracy assessment. While it is true that scales for the interpretation of the kappa coefficient exist, with that provided by Landis and Koch (1977) widely used in remote sensing, there are substantial problems in their use. For example, there are a range of scales available (e.g. Figure 3) with no obvious way to choose between them and a scale could readily be constructed for other indices such as overall accuracy. More critically, it should be readily apparent that such interpretation scales are arbitrary and cannot be of universal applicability (Sim and Wright, 2005; Vach, 2005; Banerjee et al., 1999). Indeed, Landis and Koch (1977) explicitly note the arbitrary nature of the scale that they proposed in their study. Some studies may, for example, require very high quality labelling and hence the thresholds dividing the scale should be set at higher values. The arbitrary and subjective nature of the scales limit their value as a means to interpret a kappa coefficient. The problems also mean that the existence of an interpretation scale does not address the inability to define a meaningful target value if using the kappa coefficient as the index of accuracy.

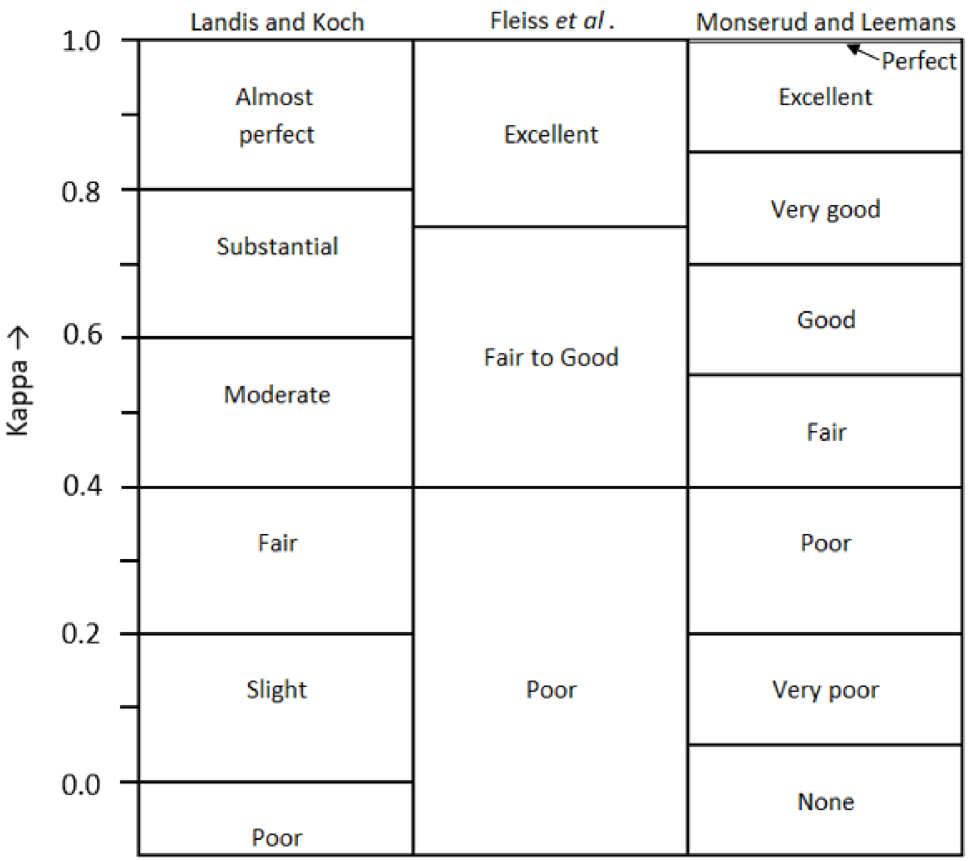

Figure 3. Three scales for the interpretation of the kappa coefficient (adapted and updated from Czaplewski, 1994). The scales are those provided by Landis and Koch (1977, page 165); Fleiss et al. (2013, page 604) and Monserud and Leemans (1992, page 285). Note that the full scale of measurement does extend to -1 but the focus is usually on positive values only.

The interpretation of a kappa coefficient can be challenging, especially if not accompanied by the confusion matrix and details of the sample of cases used in its estimation. Indeed it is widely suggested that that the provision of a kappa coefficient alone is misleading and that per-class measures and/or indices of bias and prevalence should accompany it (Byrt et al., 1993; Lantz and Nebenzahl, 1996; Cicchetti and Feinstein, 1990); the provision of the confusion matrix and details of the sample used in its construction would also help as they can provide the additional information needed to interpret a kappa coefficient. A variety of challenges is encountered in interpreting the magnitude of a kappa coefficient. In particular, two paradoxes commonly arise (Feinstein and

Cicchetti, 1990; Lantz and Nebenzahl, 1996; Hoehler, 2000; Sim and Wright, 2005). First, there is the situation in which there may be high level of agreement indicated by p_o but a low kappa coefficient. Second, unbalanced matrix marginal values can help produce a high kappa coefficient, especially if the marginals are asymmetrically imbalanced (Feinstein and Cichetti, 1990). These paradoxes arise because the estimation of the kappa coefficient is influenced by prevalence and bias between the raters (Byrt et al., 1993; Lantz and Nebenzahl, 1996; Hoehler, 2000). Both paradoxes can be explained by the distribution of cases within the confusion matrix. The first paradox arises because of the effect of prevalence on the estimation of the kappa coefficient and is positively related to the difference between *a* and *d* (Figure 1). The second paradox is related to bias effects that occur when the two sources of class labels used to form the confusion matrix differ in the proportion of cases with the trait of interest and varies as a function of the difference between *b* and *c* (Figure 1). Critically, the manner in which cases are distributed in the confusion matrix and its resulting marginal values can greatly impact on the magnitude of the kappa coefficient.

It is sometimes claimed that the whole confusion matrix is used in the estimation of the kappa coefficient. This claim, however, is untrue; the estimation of the kappa coefficient is based on the main diagonal and marginal values only (Nishii and Tanaka, 1999; Jiang and Liu, 2011). It is, for example, possible in a multi-class classification to change the entries in the matrix but maintain the same diagonal and marginal values and hence kappa coefficient. Because of the prevalence and bias effects noted above, knowledge of all of the elements of the matrix is, however, useful in interpreting a kappa coefficient (Lantz and Nebenzahl, 1996).

The factors that influence the magnitude of the kappa coefficient are well-known but the size and importance of the issues may not always be apparent. To help demonstrate problems in the interpretation and use of the kappa coefficient it may be helpful to explore some simple scenarios as examples. As a starting point, a range of possible values for the kappa coefficient can be obtained for any given level of agreement (p_o). This range can be explored by moving cases around the confusion matrix in a manner that maintains the proportion of correct agreement. The maximum and minimum kappa coefficient possible may also be estimated given an understanding of how the distribution of cases in a confusion matrix impacts on the estimation of the kappa coefficient (Lantz and Nebanzahl, 1996). Figure 4 shows the relationship between the maximum and minimum kappa coefficient values that can be obtained for all possible proportions of correct agreement. A key feature to note is the extremely large difference between the maximum and minimum kappa coefficient at each value for the proportion of correct agreement. For example, with the very high level of agreement of p_o=0.95 it would be perfectly possible for a kappa coefficient of between −0.026 and 0.900 to be estimated. Moreover, this very wide range of possible values for the kappa coefficient covers every single level of the widely used interpretation scale of Landis and Koch (1977). Thus, with 95% of the cases correctly labelled the use of the kappa coefficient could result in the level of agreement interpreted as being anything from poor to almost perfect inclusive (Figure 3).

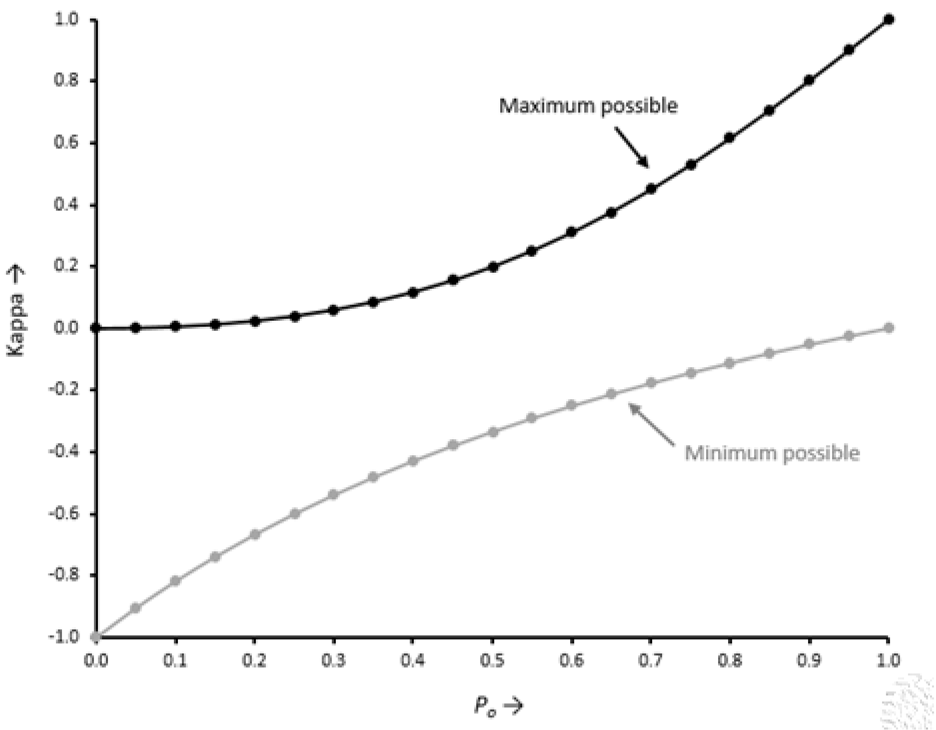

Figure 4. Relationships between the maximum and minimum possible kappa coefficient with overall accuracy (p_o).

The confusion matrices for the extreme values of the kappa coefficient when p_o=0.95 are shown in Figure 5 and highlight the effect of bias on the maximum value and prevalence on the minimum value. Importantly, very different interpretations of classification accuracy could be drawn from the use of the kappa coefficient and overall accuracy. Even though 95% of the cases in the confusion matrix have been correctly labelled it would be possible for a negative kappa coefficient to be estimated that would indicate the level of agreement was less than that due to chance. While the minimum kappa coefficient could be usefully interpreted as highlighting a poor classification, with virtually all cases allocated to one class and the accuracy for one class zero, intermediate values could be obtained. For example, Figure 6 shows one matrix for which the overall accuracy and producer's accuracy for each class are all approximately 95%, highlighting a very accurate classification. The kappa coefficient for the matrix in Figure 6 is 0.592 which lies in the range of 'moderate' agreement in the Landis and Koch (1977) scale yet the classification meets an exacting Anderson-type target of an overall accuracy of 95% with a producer's accuracy of at least 95% for each class; note purely for ease of argument the focus is on the accuracy estimate itself relative to the target value and not its associated confidence interval although the use of the latter may sometimes be appropriate.

475	50	525
0	475	475
475	525	1000

(a)

0	25	25
25	950	975
25	975	1000

(b)

Figure 5. Example confusion matrices to illustrate the range of possible kappa coefficients that could arise for a classification with p_o=95% (Figure 4). The layout of the matrices is as defined in Figure 1 and a sample of 1000 cases assumed. (a) Matrix for the maximum possible kappa coefficient, κ = 0.900 (95% CI 0.873 - 0.927). (b) Matrix for the minimum possible kappa coefficient, κ = -0.026 (95% CI -0.033 - -0.019).

40	47	87
2	911	913
42	958	1000

Figure 6. Confusion matrix for a classification that meets an Anderson-type target of an overall accuracy ≥ 95% and the producer's accuracy for each class are approximately equal and ≥95%. For this matrix, p_o = 95.1%, and the producer's accuracies are 95.23% and 95.09%. The kappa coefficient for this matrix is κ = 0.592 (95% CI 0.496 – 0.698).

A key concern with the use of the kappa coefficient is its prevalence dependency (Byrt et al., 1993; Feinstein and Cicchetti, 1990; Sim and Wright, 2005). Again, while this is well-known it may be that the size of the effect is not fully appreciated. Figure 7 shows how the magnitude of the kappa coefficient varies with prevalence for three scenarios with a fixed overall accuracy (Vach, 2005): overall accuracies of 85%, 90% and 95%. Note the magnitude of the kappa coefficient varies greatly and the effects of prevalence are especially apparent at very large or low values of prevalence. In addition, a single value for the kappa coefficient could be associated with classifications of different overall accuracy due to differences in prevalence. Indeed differences in prevalence could change the apparent order or ranking of a series of classifications. For example, a classification could be viewed as being more accurate than another in terms of overall accuracy yet the exact opposite trend could be provided by the kappa coefficients; ranking classifications in terms of accuracy requires careful interpretation. The effect of prevalence variations is also very large and is further illustrated in Figure 8 which shows matrices for four scenarios in which the overall accuracy and producer's accuracy for each class are fixed at 90% but which differ in prevalence. Each of the four matrices shown in Figure 8 have the same overall accuracy and producer's accuracies but the magnitude of the kappa coefficient differs greatly. Indeed the 95% confidence intervals fitted to the four estimates of the kappa coefficient only just touch for two of the scenarios shown (Figure 8b and 8c). Comparing kappa coefficients is, therefore, challenging if there are differences in prevalence. Thus, the kappa coefficient would not be a suitable measure if comparing classifications of study areas that may contain the same classes but at different abundances; similar problems with prevalence dependency may be observed with many other measures of accuracy. Would a difference in the magnitude of observed kappa coefficients indicate a difference in the quality of class labelling or merely reflect the variations in class prevalence?

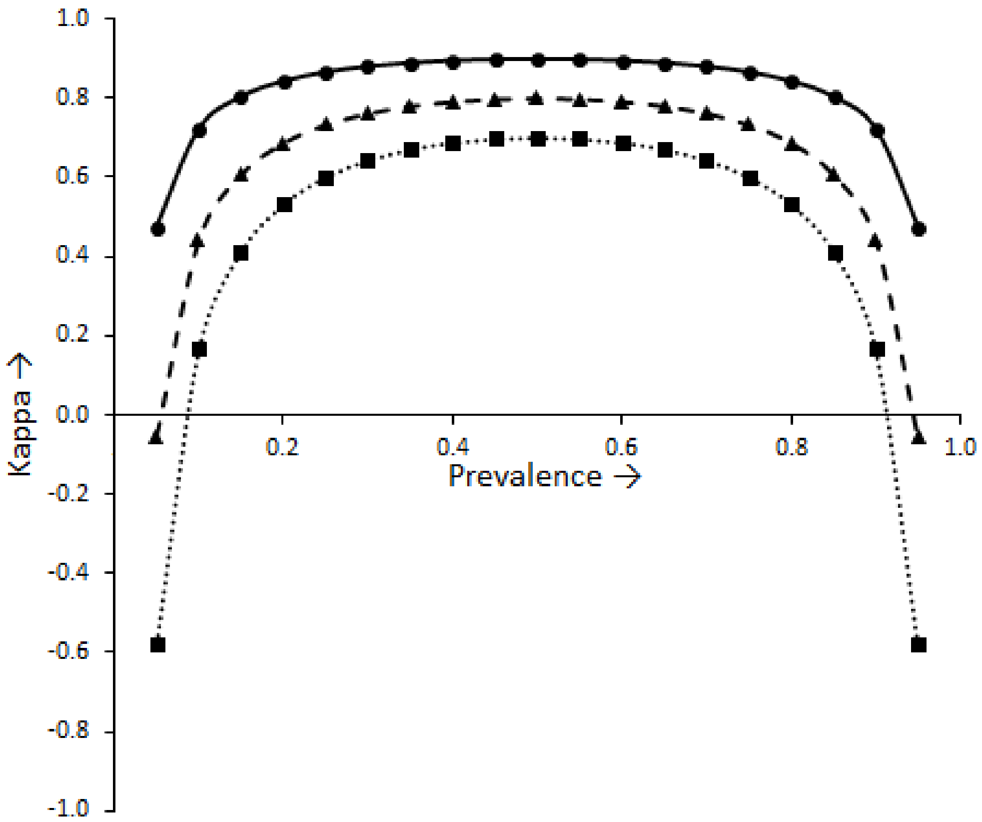

Figure 7. Variation in the magnitude of the kappa coefficient with prevalence for three fixed value of overall accuracy. Three scenarios are shown in which the marginal values (i.e. $n_{\cdot+}$ and $n_{+\cdot}$) are equal and the overall accuracy is 85% (dotted line with square symbols), 90% (dashed line with triangular symbols) and 95% (solid line with circular symbols).

450	50	500
50	450	500
500	500	1000

Prevalence = 0.50
κ = 0.800 (95% CI 0.763 - 0.837)

(a)

90	90	180
10	810	820
100	900	1000

Prevalence = 0.10
κ = 0.590 (95% CI 0.520 - 0.661)

(b)

45	95	140
5	855	860
50	950	1000

Prevalence = 0.05
κ = 0.432 (95% CI 0.344 - 0.520)

(c)

9	99	108
1	891	892
10	990	1000

Prevalence = 0.01
κ = 0.137 (95% CI 0.055 - 0.218)

(d)

Figure 8. Confusion matrices for a scenario in which there is constant agreement on an overall and per-class basis (p_o = 0.9, producer's accuracy for each class = 90%) but varying prevalence. (a) prevalence = 0.5 (i.e. the two classes have equal abundance), (b) prevalence = 0.10, (c) prevalence = 0.05, and (d) prevalence = 0.01.

The various problems associated with the interpretation of the kappa coefficient make comparison of kappa coefficients difficult, especially if the comparison is between studies of regions of dissimilar prevalence (Ubserax, 1987; Byrt et al., 1993; Vach, 2005; Sim and Wright, 2005). A major concern is that the magnitude of a kappa coefficient and its possible range of values reflect the nature of the population being studied (e.g. prevalence) (Byrt et al., 1993; Lantz and Nebenzahl, 1996). The kappa coefficient has been widely promoted as a summary statistic that is meant to convey information on thematic accuracy but it is a poor tool as it is highly mis-leading (Maclure and Willett, 1987). The kappa coefficient is not well suited for use in accuracy assessment. Rather than use the kappa coefficient because other studies have done so, and perpetuate a mistake, researchers should select an accuracy measure appropriate for the task in-hand recognising that different measures of accuracy reflect different aspects of quality and may require careful interpretation. Inspired by the comments of the referees on this article, as part of an effective peer review process, referees and editors should perhaps challenge the use of a measure such as the kappa coefficient in applications such as accuracy assessment and comparison for which it is unsuitable.

Finally on the issue of prevalence, it may be worth remembering that at the outset one key reason for not using overall accuracy was because of its sensitivity to the effect of variations in prevalence. This dependency is well known with $p_o = (\theta P_+ + (1\text{-}\theta)P_-)$. Overall accuracy is certainly an imperfect measure, as is any omnibus index (Byrt et al., 1993; Cicchetti and Feinstein, 1990), and no single measure will be universally ideal for accuracy assessment (Stehman, 1997a) but the kappa coefficient does not solve the problems associated with overall accuracy. That the kappa coefficient is prevalent dependent should come as no surprise given it is calculation from p_o and p_e in equation 1. Kappa is simply a rescaled version of p_o and p_e is prevalent dependent as prevalence is included in its calculation (equation 3). Because of the limitations of overall accuracy researchers have been encouraged to state per-class accuracies, such as user's and producer's accuracy, in addition (e.g. Liu et al., 2007; Stehman, 2000; Olofsson et al., 2014). A further enhancement would be to follow further good practices such as the provision of the confusion matrix and details of the sample used in its construction to allow estimation of other measures, even the kappa coefficient, if desired (Olofsson et al., 2013, 2014). It is difficult to identify how the provision of the kappa coefficient adds positively to this situation. The kappa coefficient alone is mis-leading so other information, notably on bias and prevalence, needs to be provided with it. The provision of a difficult to interpret measure such as the kappa coefficient that must be accompanied by additional measures such as bias and prevalence to aid interpretation does not help communicate accuracy information in a clear and succinct way. Then, in addition, there are concerns about the way chance is modelled and used. Given that the kappa coefficient is estimated from overall accuracy, it is evident that the estimation of the kappa coefficient is an unhelpful and unnecessary step in the assessment or comparison of classification accuracy.

4. Conclusions

The kappa coefficient is widely promoted and used as a measure of thematic accuracy in remote sensing. The publications that promoted the use of the kappa coefficient have played an enormously influential role to inspire thought concerning rigorous quantitative assessments of classifications but promoted an inappropriate index. The reasons espoused for the use of the kappa coefficient are flawed and/or irrelevant as they apply equally well to other measures. Critically, the kappa coefficient is not an index of accuracy but a measure of the level of agreement observed beyond chance that is obtained using a model of chance that is inappropriate to the typical accuracy assessment scenario. Not only is the effect of chance agreement mis-estimated it is, however, irrelevant to an accuracy assessment which seeks to indicate the amount of error, and thereby correctness, in the labelling with the source of error inconsequential. The kappa coefficient is an inappropriate index to use to describe classification accuracy.

Many of the concerns with the kappa coefficient have been known for decades and it may be that its continued use in remote sensing is, in part, because the problems are viewed as being small and insubstantial. Here, emphasis has been placed on indicating the size and nature of the problems with the kappa coefficient by showing how its magnitude can vary as a function of basic properties of a study such as prevalence. Critically, simple examples have been used to show the unsuitability of the kappa coefficient for the description of accuracy and its comparison. For example, it was shown that classifications with an overall accuracy of 95% could have a kappa coefficient that lay within the range from -0.026 to 0.900. The difficulty of interpreting the estimated kappa coefficients is further highlighted by noting that the entire spread of possible values covers the complete range of the widely used Landis and Koch (1977) interpretation scale. Furthermore, if the classification satisfied a demanding Anderson-type target that required the producer's accuracy for each class be ≥95% the kappa coefficient for this very accurate classification would be interpreted as showing only moderate agreement. A key problem is the effect of variations in class abundance or prevalence, the very problem highlighted in criticisms of overall accuracy. Differences in prevalence make the comparison of kappa coefficients very difficult, a researcher will be unsure if a difference reflects dissimilarity in the level of agreement or of the populations being studied. Overall accuracy on the other hand, while flawed, does have a clear meaning and, relative to kappa, is simple to estimate.

Different measures of accuracy reflect different aspects of a classification (Hand, 2012). Care must, therefore, be taken to ensure that a measure of accuracy that is appropriate for the task in-hand is adopted. There are many possible motivations and interests in an accuracy assessment which makes the provision of universal recommendations difficult. The literature on accuracy assessment can at times be challenging and other researchers may be better qualified to comment with authority and clarity on the topic but the common practice of using the kappa coefficient to indicate classification accuracy is flawed. Indeed, from the discussion above it is recommended that the kappa coefficient be dropped from the community's toolbox or at least used only sparingly and when good reason for its estimation exists such as in the assessment of agreement in class labelling among multiple interpreters. Although there are sometimes challenges to fully documenting an accuracy assessment, the provision of overall accuracy and per-class accuracy values together with the confusion matrix, set in the context of broader good practices (e.g. Olofsson et al., 2014; Stehman and Foody, 2019), should meet the objectives of most accuracy assessments. The provision of such information also allows assessments from other perspectives and the estimation of other measures, including even the kappa coefficient if desired, in order to meet the specific aims of a study. Comparisons of accuracy statements can be undertaken using overall accuracy and per-class accuracy using the same approach suggested for kappa if the samples involved are independent. If the samples are not independent, as is often the case in remote sensing research, alternative means to compare classification accuracy such as the McNemar test may be used. The kappa coefficient does not add positively to such accuracy assessments and comparisons. Given the challenges with its interpretation, the kappa coefficient should, therefore, not be used and reported routinely.

Acknowledgments

This article draws on experiences in accuracy assessment over many years that have benefited from inputs by colleagues ranging from article referees through to research collaborators which is much appreciated. The highly constructive and thoughtful comments from the three referees and editors are greatly appreciated and helped enhance the final version of the article.

References

Allouche, O., Tsoar, A. and Kadmon, R., 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, 43(6), 1223-1232.

Anderson, J. R., 1971. Land-use classification schemes, *Photogrammetric Engineering*, 37, 379-387.

Anderson, J. R., Hardy, E. E., Roach, J. T. and Witmer, R. E., 1976. *A Land Use and Land Cover Classification System for Use with Remote Sensor Data*, Geological Survey Professional Paper 964, 28pp.

Ariza-López, F.J., Rodríguez-Avi, J., Alba-Fernández, M.V. and García-Balboa, J.L., 2019. Thematic accuracy quality control by means of a set of multinomials. *Applied Sciences*, 9, 4240.

Banerjee, M., Capozzoli, M., McSweeney, L. and Sinha, D., 1999. Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, 27(1), 3-23.

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

Byrt, T., Bishop, J. and Carlin, J.B., 1993. Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46(5), 423-429.

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

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

Comber, A., Fisher, P., Brunsdon, C. and Khmag, A., 2012. Spatial analysis of remote sensing image classification accuracy. *Remote Sensing of Environment*, 127, 237-246.

Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, 37(1), 35-46.

Congalton, R.G. and Green, K., 2009. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. Second edition, CRC press, Boca Raton.

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(1), 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(12), 1671-1678.

Czaplewski, R. L. 1994. *Variance Approximations for Assessments of Classification Accuracy*. Res. Pap. RM-316. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Forest and Range Experiment Station. 29 p.

Donner, A., Shoukri, M.M., Klar, N. and Bartfay, E., 2000. Testing the equality of two dependent kappa statistics. *Statistics in Medicine*, 19(3), 373-387.

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

Fielding, A.H. and Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24(1), 38-49.

Finn, J.T., 1993. Use of the average mutual information index in evaluating classification error and consistency. *International Journal of Geographical Information Science*, 7(4), 349-366.

Fitzgerald, R.W. and Lees, B.G., 1994. Assessing the classification accuracy of multisource remote sensing data. *Remote Sensing of Environment*, 47(3), 362-368.

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

Fleiss, J.L., Levin, B. and Paik, M.C., 2013. *Statistical Methods for Rates and Proportions*. Third edition, John Wiley & Sons.

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

Foody, G.M., 2004. Thematic map comparison: evaluating the statistical significance of differences in classification accuracy. *Photogrammetric Engineering and Remote Sensing*, 70(5), 627-633.

Foody, G.M., 2008. Harshness in image classification accuracy assessment. *International Journal of Remote Sensing*, 29(11), 3137-3158.

Foody, G.M., 2009. Classification accuracy comparison: hypothesis tests and the use of confidence intervals in evaluations of difference, equivalence and non-inferiority. *Remote Sensing of Environment*, 113(8), 1658-1663.

Foody, G.M., 2010. Assessing the accuracy of land cover change with imperfect ground reference data. *Remote Sensing of Environment*, 114(10), 2271-2285.

Foody, G.M., 2011. Latent class modeling for site-and non-site-specific classification accuracy assessment without ground data. *IEEE Transactions on Geoscience and Remote Sensing*, 50(7), 2827-2838.

Hand, D. J., 2012. Assessing the performance of classification methods, *International Statistical Review*, 80 (3), 400-414.

Hoehler, F.K., 2000. Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *Journal of Clinical Epidemiology*, 53(5), 499-503.

Hudson, W. D. and Ramm, C. W., 1987. Correct formulation of the Kappa coefficient of agreement, *Photogrammetric Engineering and Remote Sensing*, 53, 421-422.

Jansen, L L F, and van der Wel, F.J., 1994. Accuracy assessment of satellite derived landcover data: A review. *Photogrammetric Engineering and Remote Sensing*, 60(4), 479-426.

Jiang, S. and Liu, D., 2011. On chance-adjusted measures for accuracy assessment in remote sensing image classification. In *ASPRS Annual Conference,* ASPRS 2011 Annual Conference Milwaukee, Wisconsin,May 1-5, 2011

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

Lantz, C.A. and Nebenzahl, E., 1996. Behavior and interpretation of the κ statistic: Resolution of the two paradoxes. *Journal of Clinical Epidemiology*, 49(4), 431-434.

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

Maclure, M. and Willett, W.C., 1987. Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology*, 126(2), 161-169.

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

Monserud, R.A. and Leemans, R., 1992. Comparing global vegetation maps with the Kappa statistic. *Ecological Modelling*, 62(4), 275-293.

Morales-Barquero, L., Lyons, M.B., Phinn, S.R. and Roelfsema, C.M., 2019. Trends in remote sensing accuracy assessment approaches in the context of natural resources. *Remote Sensing*, 11, 2305.

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

Olofsson, P., Foody, G.M., Stehman, S.V. and Woodcock, C.E., 2013. Making better use of accuracy data in land change studies: estimating accuracy and area and quantifying uncertainty using stratified estimation. *Remote Sensing of Environment*, 129, 122-131.

Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E. and Wulder, M.A., 2014. Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, 148, 42-57.

Pontius Jr, R.G., 2000. Comparison of categorical maps. *Photogrammetric Engineering and Remote Sensing*, 66(8), 1011-1016.

Pontius Jr, R.G. and Millones, M., 2011. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, 32(15), 4407-4429.

Pontius, R.G. and Parmentier, B., 2014. Recommendations for using the relative operating characteristic (ROC). *Landscape Ecology*, 29(3), 367-382.

Rogan, W. J. and Gladen, B. (1978) Estimating prevalence from the results of a screening test, American Journal of Epidemiology, 107, 71-76.

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(2), 223-227.

Sim, J. and Wright, C.C., 2005. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy*, 85(3), 257-268.

Smits, P.C., Dellepiane, S.G. and Schowengerdt, R.A., 1999. Quality assessment of image classification algorithms for land-cover mapping: A review and a proposal for a cost-based approach. *International Journal of Remote Sensing*, 20(8), 1461-1486.

Stehman, S., 1996. Estimating the kappa coefficient and its variance under stratified random sampling. *Photogrammetric Engineering and Remote Sensing*, 62(4), 401-407.

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

Stehman, S.V., 1997b. Estimating standard errors of accuracy assessment statistics under cluster sampling. *Remote Sensing of Environment*, 60(3), 258-269.

Stehman, S.V., 2000. Practical implications of design-based sampling inference for thematic map accuracy assessment. *Remote Sensing of Environment*, 72(1), 35-45.

Stehman, S.V. and Foody, G.M., 2009. Accuracy assessment. In: Warner, T A, Nellis, N. D. and Foody, G M.(editors) *The SAGE Handbook of Remote Sensing*, Sage, London, 297-309.

Stehman, S.V. and Foody, G.M., 2019. Key issues in rigorous accuracy assessment of land cover products. *Remote Sensing of Environment*, 231, 111199.

Story, M. and Congalton, R.G., 1986. Accuracy assessment: a user's perspective. *Photogrammetric Engineering and Remote Sensing*, 52(3), 397-399.

Thron, C. and Miller, V., 2015. Persistent confusions about hypothesis testing in the social sciences. *Social Sciences,* 4(2), 361-372.

Tsutsumida, N. and Comber, A.J., 2015. Measures of spatio-temporal accuracy for time series land cover data. *International Journal of Applied Earth Observation and Geoinformation*, 41, 46-55.

Türk, G., 1979. GT index: A measure of the success of prediction. *Remote Sensing of Environment*, 8(1), 65-75.

Türk, G. 2002. Map evaluation and 'chance correction'. *Photogrammetric Engineering and Remote Sensing*, 68, 123–133.

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

Vach, W., 2005. The dependence of Cohen's kappa on the prevalence does not matter. *Journal of Clinical Epidemiology*, 58(7), 655-661.

Wu, S.M., Whiteside, U. and Neighbors, C., 2007. Differences in inter-rater reliability and accuracy for a treatment adherence scale. *Cognitive Behaviour Therapy*, 36(4), 230-239.

Ye, S., Pontius Jr, R.G. and Rakshit, R., 2018. A review of accuracy assessment for object-based image analysis: From per-pixel to per-polygon approaches. *ISPRS Journal of Photogrammetry and Remote Sensing*, 141, 137-147.