

The evaluation and comparison of thematic maps derived from remote sensing

Giles M. Foody

School of Geography, University of Southampton,
Southampton, S017 1BJ, UK
Tel: +44 (0)2380 595 493; Fax: +44 (0)2380 593 295
g.m.foody@soton.ac.uk

Abstract

The accuracy of thematic maps derived from remote sensing is often viewed negatively. This reflects difficulties in mapping but also issues connected with accuracy evaluation targets and assessment approaches. This paper focuses on the latter issues suggesting that the widely used target accuracy of 85% may often be inappropriate and that the approach to accuracy assessment adopted commonly in remote sensing can be pessimistically biased. Problems are also encountered in the comparison of thematic maps, hindering research that seeks to increase classification accuracy which is often based upon evaluations of the accuracy of maps derived from a set of classification algorithms. It is hoped that a greater awareness of the problems encountered in accuracy assessment and comparison may help ensure that perceptions of classification accuracy are realistic and fair.

Keywords: accuracy, classification, difference, target, imagery

1 Introduction

Thematic mapping via an image classification is one of the most commonly undertaken analyses of remotely sensed data. The accuracy of the classification is clearly a key factor in determining its fitness for purpose and is the main criterion used in evaluating a thematic map. The comparison of accuracy statements is also the basis of evaluations of different classifiers. Although the importance of accuracy assessment is widely acknowledged there are, however, many problems in map evaluation and comparison (Pontius, 2000, 2002; Foody, 2002; Pontius and Cheuk, 2006; Wulder *et al.*, 2006).

The accuracy of image classifications used in thematic mapping is often perceived as being inadequate for many users (Townshend, 1992; Wilkinson, 1996). This has driven considerable research aimed at increasing map accuracy. However, no upward trend in accuracy has been observed with a mean accuracy, expressed as a kappa coefficient of agreement, of ~0.66 (Wilkinson, 2005). It is, therefore, unsurprising that users question the accuracy of thematic maps derived from remote sensing. Sometimes this questioning arises when a thematic map is used for an application other than that it was designed for. For example, this may occur when a thematic map developed for specific small cartographic scale applications is used at much larger scales than it was intended for. There is additionally considerable anecdotal evidence of users questioning the accuracy of maps, often on the basis of very localized assessments. Moreover, when comparing maps, especially in the evaluation of classification algorithms, differences in accuracy may not be fully appreciated. In this paper it is suggested that the assessment, interpretation and comparison of classification accuracy is often difficult and aspects may often be made from an overly harsh perspective.

2 Accuracy target

The evaluation of the quality of a thematic map should ideally be based on a set of criteria defined in advance of its production. The criteria adopted may be expected to vary as a function of variables such as the nature of the remotely sensed imagery used (e.g. spatial and spectral resolution), the classes defined (e.g. number and detail of classes) and user needs (e.g. tolerance to error). There are, therefore, no universally defined accuracy standards for thematic mapping from remote sensing (e.g. Loveland *et al.*, 1999; Kerr and Cihlar, 2004). However, concern is typically focused on map accuracy and the definition of a minimum level of accuracy required provides a basic criterion for evaluative purposes. Maps are typically evaluated in relation to the magnitude of their estimated accuracy. Although there is no universal standard target accuracy to achieve and a target value is often not stated explicitly, one value that has been widely used as a target in thematic mapping via an image classification is to achieve an accuracy of $\geq 85\%$ correct allocation. This 85% accuracy is often viewed explicitly as the standard of acceptability for thematic mapping from remotely sensed imagery (e.g. Wright and Morrice, 1997; Abeyta and Franklin, 1998; Brown *et al.*, 2000; Treitz and Rogan, 2004).

The 85% target accuracy often seems to be used unquestioningly and adopted as a universal standard for thematic mapping in remote sensing (e.g. Fisher and Langford, 1996; Weng, 2002; Rogan *et al.*, 2003; Bektas and Goksel, 2004; Wulder *et al.*, 2006). Thus, the 85% target has been used in studies spanning a vast range of diverse applications including global mapping of broad land cover classes (Scepan, 1999), the mapping of very detailed classes at a very local scale (McCormick, 1999) and assessments of change detection (Sader *et al.*, 2001).

The origin of this 85% target accuracy can often be traced back to the influential work of Anderson *et al.* (1976). However, Anderson *et al.* (1976) do not discuss the issue in detail or set out to propose a universally adoptable set of map evaluation criteria. Instead Anderson *et al.* (1976) focus on the classification schemes that could be used with remotely sensed data. The evaluation of the accuracy of the derived classifications, although clearly an important issue, was not a major issue and was only discussed briefly and tentatively in the article, aware that the sensing technology was rapidly developing and that it is unlikely that there is one ideal approach to promote. Furthermore, Anderson *et al.* (1976) were explicit in relation to the nature of the thematic map under study and justify the adoption of the 85% value which was specified for a particular application scenario. The scenario discussed by Anderson *et al.* (1976) was the mapping of broad land cover classes, such as those at Anderson level I, at scales in the range of 1:250,000 to 1:2,500,000. The basis of the 85% target in this mapping was because this would be comparable to the accuracy of land cover maps derived from aerial photograph interpretation undertaken previously in work associated with the USDA's Census of Agriculture. That is, an aim was to emulate the accuracy that could be achieved for a very specific task through the application of conventional approaches such as aerial photograph interpretation. Thus the 85% target accuracy was specified by Anderson *et al.* (1976) for mapping broad land cover classes (Anderson level I, 9 broad classes) from Landsat 1 sensor data (e.g. Landsat MSS imagery with ~ 80 m spatial resolution and acquired in 4 spectral wavebands). The suggested thematic map evaluation criteria were not proposed as being universally applicable. They were not, for instance, suggested for detailed class mapping of local regions from imagery of the type available from contemporary satellite sensing systems. The Anderson *et al.* (1976) proposal was made in relation to mapping a small number of classes from, what may be considered in this context to be, fine spatial resolution multispectral

imagery. One would expect that the accuracy with which a region could be mapped would vary as a function of issues such as the number and detail of the classes and nature of the sensor (e.g. its spectral and spatial resolution). Many studies, for example, aim to map detailed classes at a large cartographic scale (Wilkinson, 2005). Such classes and scales were explicitly outside the scope of discussion of Anderson *et al.* (1976) who suggested that such mapping would require large amounts of ancillary information. Despite this, much of the remote sensing community appears to have latched on to the 85% target accuracy as some general criterion to apply, irrespective of the specific nature of the mapping task in-hand. Additionally, the community of map users seems to have followed suit and appear to have adopted the 85% target too. It is unclear why the 85% target has been used so widely, especially as it may not be realistic. If, for example, the aim is to map a small number of very spectrally separable classes then the target should perhaps be set at a higher value. Alternatively, and perhaps more commonly, if there are many classes that are only subtly different it seems reasonable to ask if the target accuracy is too high and unachievable. To be of value, a target should really be specified for the particular application in-hand and be realistic. It may be that 85% is often a perfectly reasonable target to adopt but it should not be accepted and used without question. In particular, it is worth stressing that Anderson *et al.* (1976) specifically report that the detailed classification schemes, such as those at Anderson levels III and IV, were beyond the scope of their discussion. This is important as, generally, an increase in the detail of the classes would be associated with a reduction in classification accuracy (e.g. Stehman *et al.*, 2003). For many mapping applications, 85% may be an unrealistically high target to set. In reality there is no single accuracy target that will be universally appropriate. In many instances the 85% target value suggested by Anderson *et al.* (1976) may be inappropriate as it may be unrealistically high for the application. Critically, however, the widely used target of 85% should not automatically be used as a criterion for the evaluation image classifications (Laba *et al.*, 2002).

One issue in the evaluation of thematic map accuracy on which the remote sensing community could, however, adopt a harsher approach is in deciding whether a thematic map produced by an image classification satisfies the target value specified. Typically, the basis of assessing the acceptability of a map is to calculate a measure of the map's accuracy and compare this directly against the target value (e.g. Hayes and Sader, 2001). The thematic map is typically judged to be sufficiently accurate if the calculated accuracy value equals or exceeds the target. However, the accuracy statement derived in most studies is just an estimate of the accuracy of the classification. In many instances it would be more appropriate to fit confidence limits to the estimate (Thomas and Allcock, 1984) and consider these when evaluating the map and deciding if the target accuracy has been achieved. Thus, for example, the accuracy statement for an image classification should perhaps ideally take the form of the estimated value \pm the half width of the confidence interval at some specified level of statistical confidence. This is important as sometimes the estimated accuracy of a classification may exceed the 85% target value but the confidence limits may suggest that it would be unwise to assume that this means the classification has achieved the target level desired. However, a classification with an estimated accuracy that barely exceeds the target value specified is often viewed as being of acceptable quality (e.g. Hayes and Sader, 2001). The casual comparison of the accuracy estimate directly against the target can give an inappropriate basis for evaluating a classification. The confidence limits fitted around the estimated value of map accuracy provide important information that should influence the evaluation of the accuracy of the map and its suitability for later application. Map producers should, therefore, be encouraged to fit

confidence limits to classification accuracy statements and promote their use in evaluating the classification's fitness for its intended application.

3 Accuracy assessment methods

Thematic map accuracy is most commonly estimated through the use of site-specific techniques based on the analysis of the entries in a confusion or error matrix (Congalton and Green, 1999; Foody, 2002). This matrix should provide a simple summary of classification accuracy and highlight the two types of thematic error that may occur: omission and commission. In reality, however, there are many problems associated with the confusion matrix and the measures of accuracy derived from it. The meaning of basic summary measures of accuracy such as the proportion of correctly allocated cases, the most widely used index of classification accuracy, is, for example, a function of the sample design used in the acquisition of the testing set (Stehman, 1995). Thus, the estimates of accuracy derived from confusion matrices constructed from testing sets drawn by simple random and stratified random sampling, without any allowance for the difference in the sample design, may differ substantially if the classes vary in abundance and spectral separability.

Many other problems are encountered with the confusion matrix, notably issues connected with assumptions that the pixels are pure and the testing set is perfectly co-located with the map. Assumptions such as these are rarely satisfied. The proportion of mixed pixels in an image is a function of the spatial resolution of the imagery and the land cover mosaic but is often very large and these pixels cannot be accommodated directly in the basic confusion matrix resulting in error. Similarly, a large amount of the error depicted in a confusion matrix may have arisen through mis-location of data points in the thematic map and in the reference data. The latter data may also contain significant uncertainty and error (Joria and Jorgenson, 1996; Khorram, 1999; Mas, 2004; Comber *et al.*, 2005). However, all disagreements between class labels in the thematic map derived from remotely sensed data and the reference data are typically interpreted, unfairly, as errors in the classification used to produce the thematic map (Fitzgerald and Lees, 1994; Foody, 2002) leading to pessimistic bias in accuracy assessment. Thus, not only may the target accuracy be unrealistically high (section 2 above), but the approach to assess accuracy may yield an unfairly negative view of thematic map accuracy. It is, of course, possible to adjust the standard approach to accuracy assessment to reduce some of the problems. For example, rather than rigidly adopt the site-specific comparison the accuracy assessment could perhaps be based on the modal class observed within the area of a defined filter window (Stehman *et al.*, 2003) or use made of modified accuracy measures that attempt to provide a degree of tolerance for mis-location (e.g. Hagen, 2003). It is important, however, to avoid the potential to optimistically bias the accuracy assessment.

The various problems of forming and interpreting the confusion matrix are often compounded by the use of inappropriate measures to quantify thematic map accuracy. There have been, for example, many calls for the remote sensing community to adopt measures such as the kappa coefficient of agreement in the assessment of classification accuracy (Congalton *et al.*, 1983; Congalton and Green, 1999; Smits *et al.*, 1999; Wilkinson, 2005). The kappa coefficient of agreement for a thematic map is based on the comparison of the predicted and actual class labels for each case in the testing set acquired for the assessment of map accuracy and may be calculated from,

$$\hat{\kappa} = \frac{p_o - p_c}{1 - p_c} \quad (1)$$

where p_o is the proportion of cases in agreement (i.e., correctly allocated) and p_c is the proportion of agreement that is expected to have arisen by chance. The arguments made for the adoption of the kappa coefficient are commonly based on statements such as its calculation corrects for chance agreement, that the entire confusion matrix is used in its derivation, that a variance term can be calculated for it which facilitates statistical comparisons and because scales exist to aid interpretation (e.g. Congalton *et al.*, 1983; Monserud and Leemans, 1992; Janssen and van der Wel, 1994; Smits *et al.*, 1999; Wheeler and Alan, 2002). The use of the kappa coefficient for map accuracy assessment has, however, often been questioned (Stehman, 1997; Turk, 2002; Jung, 2003). Indeed, each of the commonly argued reasons for using the kappa coefficient as a measure of map accuracy is open to criticism. For example, some of the arguments made for the adoption of the kappa coefficient are incorrect. The kappa coefficient is not, for instance, calculated from the entire matrix but on the basis of its main diagonal and marginals (Stehman, 1997; Nishii and Tanaka, 1999). Some of the arguments used for the adoption of the kappa coefficient fail to recognise that they apply equally to other measures of map accuracy. For example, a variance term can be derived for many other measures of accuracy that are widely used, including standard statements based on the percentage of correctly allocated cases, and be used in evaluating the statistical significance of differences in classification accuracy (Foody, 2004); this is an issue discussed further in section 4 below. Additionally, widely used scales to interpret the kappa coefficient are problematic and arbitrary (Manel *et al.*, 2001; Di Eugenio and Glass, 2004). Perhaps most importantly, the allowance for chance agreement, probably the most widely cited reason for the adoption of the kappa coefficient as a measure of map accuracy, is a problem as the degree of chance agreement may be overestimated, leading to an underestimation of map accuracy (Foody, 1992), and, more fundamentally, that chance correction is completely unnecessary (Turk, 2002). Turk (2002) argues that making some correct allocations by chance rather than by design is a windfall gain, is not something the users or producers of thematic maps should necessarily worry about. If the aim of an analysis is to state the accuracy of a thematic map derived from an image classification then the source of error is unimportant. This application requires an index of map accuracy and not of the map producing technology. One such index that may commonly be appropriate is the percentage of correctly allocated cases. There is no need to correct for chance agreement and the use of measures such as the kappa coefficient may have the effect of suggesting, on naïve inspection, that map accuracy is lower than it really is.

4 Comparison of accuracy statements

There is often a desire to compare thematic maps and in particular their accuracy. It is common, for example, to seek an evaluation of two or more classifiers. Much of this work has been driven by the perceived failure of thematic maps to meet user needs and a desire to achieve more accurate classifications. Since map accuracy statements are typically just estimates of accuracy it is important that sampling error issues are considered in their comparison.

Often assessments of the statistical significance of differences in map accuracy have based on the comparison of accuracy estimates expressed in terms of the kappa coefficient of

agreement. This situation reflects, in part, the explicit discussion of the comparison of kappa coefficients in the literature. Indeed, the ability to rigorously compare kappa coefficients has been advocated as a valuable feature since the kappa coefficient was first introduced to the remote sensing community (Congalton and Mead, 1983; Congalton *et al.*, 1983; Rosenfield and Fitzpatrick-Lins, 1986; Janssen and van der Wel, 1994; Smits *et al.*, 1999). As noted above, there are, however, concerns about the use of kappa coefficient for accuracy assessment and comparison.

The aim of the map comparison is to determine if the difference in the derived estimates can be inferred to indicate a difference in the associated population parameters of map accuracy. Drawing on an earlier article (Foody, 2004), the significance of the difference in accuracy between two maps with independent kappa coefficients, represented by $\hat{\kappa}_1$ and $\hat{\kappa}_2$, may be evaluated from,

$$z = \frac{\hat{\kappa}_1 - \hat{\kappa}_2}{\sqrt{\hat{\sigma}_{\kappa_1}^2 + \hat{\sigma}_{\kappa_2}^2}} \quad (2)$$

where $\hat{\sigma}_{\kappa_1}^2$ and $\hat{\sigma}_{\kappa_2}^2$ represent the estimated variances of the derived coefficients. The significance of the difference between the two kappa coefficients may be assessed by comparing the value of z calculated from equation 2 against tabulated values. For the simple situation of determining if there is a difference between two kappa coefficients (a two-sided test) the null hypothesis (H_0), of no significant difference, can be rejected at the widely used 5% level of significance if $|z| > 1.96$ (Congalton *et al.*, 1983; Rosenfield and Fitzpatrick-Lins, 1986; Congalton and Green, 1999). This equation has formed the basis of many map comparison analyses. However, the use of this approach may often be inappropriate. Critically, the approach based on equation 2 is explicitly based on an assumption that the two samples used are independent. Although this issue is often noted and was explicit in Cohen's (1960) paper that introduced the kappa coefficient of agreement this assumed condition is often not satisfied in remote sensing applications. This is because in many remote sensing studies the aim is to compare two or more classification approaches and the same ground data set used in the evaluation of the accuracy of each classification. Since the same sample of data is used in the derivation of each kappa coefficient, the assumption of independence is not satisfied and the approach outlined above should not be used to evaluate the statistical significance of differences in map accuracy indicated by the derived kappa coefficients (Foody, 2004; de Leeuw *et al.*, 2006). Failure to note the distinction between related and independent samples in a comparison of accuracy statements can lead to mis-interpretation of the degree of difference in accuracy that exists (Foody, 2004).

The method for comparing kappa coefficients represented by equation 2 should not be used for the comparison of kappa coefficients that were estimated with the use of the same or related samples (McKenzie *et al.*, 1996; Donner *et al.*, 2000). Alternative approaches may, however, be adopted. For example, Donner *et al.* (2000) discuss the accommodation for the dependence between kappa coefficients arising through the use of related samples that have been proposed and extend the method used for data from independent samples to account for the covariance between kappa statistics due to the use of a related sample. This approach is appropriate for the comparison of kappa coefficients derived from classifications with a dichotomous outcome

and the testing for the significance of the difference between two kappa coefficients derived from a related sample is based on,

$$z = \frac{\hat{K}_1 - \hat{K}_2}{\sqrt{(\hat{\sigma}_{\kappa_1}^2 + \hat{\sigma}_{\kappa_2}^2 - 2\hat{\sigma}_{\hat{\kappa}_1\hat{\kappa}_2})}} \quad (3)$$

where $\hat{\sigma}_{\hat{\kappa}_1\hat{\kappa}_2}$ represents the estimated covariance between \hat{K}_1 and \hat{K}_2 (Donner *et al.*, 2000).

An alternative approach to comparing kappa coefficients derived from related samples is based upon the use of resampling techniques (McKenzie *et al.*, 1996). This approach involves deriving a large number of samples from the original sample to form a probability distribution of the statistic. The method presented by McKenzie *et al.* (1996) uses Monte Carlo permutation tests in the determination of the statistical significance of the difference between two kappa coefficients derived using related samples. For this, the variable in common to the two classifications (e.g. the ‘actual’ class label derived from the reference data set) are randomly shuffled and the kappa coefficients recomputed. For each permutation, the difference between the kappa coefficients derived is estimated. The number of times that the original difference in the kappa coefficients is equalled or exceeded by the difference in the randomly permuted values derived in the analysis is noted, incremented by 1 and divided by the total number of permutation plus 1 to derive a proportion. For the common situation in which a two-sided test (H_0 that there is no significant difference between the two kappa coefficients) at the 5% level of significance is being undertaken, the difference between two kappa coefficients derived with related samples would be regarded as being statistically significant if the computed proportion was less than 0.05.

As noted above, there are, however, concerns about the use of the kappa coefficient as a measure of accuracy. There is nothing unique about the kappa coefficient in terms of the ability to statistically compare values. A variety of measures may be used to estimate map accuracy and techniques are available to allow their comparison, for situations in which the samples are independent and related. For example, it is possible to compare the proportion of cases correctly allocated, the most commonly used measure of accuracy used in remote sensing (Trodd, 1995). In the situation in which the samples are independent, there is no need to calculate the kappa coefficient and use equation 2, the significance of the difference between two proportions may be estimated instead from,

$$z = \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (4)$$

where x_1 and x_2 are the number of correctly allocated cases in two independent samples of size n_1 and n_2 respectively and $p = \frac{x_1 + x_2}{n_1 + n_2}$. The estimated proportion of correctly allocated cases

for each map (e.g. $\frac{x_1}{n_1}$) can, of course, be converted to the percentage correct allocation by

multiplying by 100 and used to indicate map accuracy. The statistical significance of the difference in accuracy between two classifications is evaluated through z in the same way as with the comparison of kappa coefficients. If, as is common in remote sensing, the samples are related the statistical significance of the difference between two proportions may be evaluated using a McNemar test (Bradley, 1968; Agresti, 1996). This test is based upon confusion matrices that are 2x2 in dimension and show the level of inter-classifier agreement in correct and incorrect allocations. The McNemar test is based on,

$$z = \frac{f_{12} - f_{21}}{\sqrt{f_{12} + f_{21}}} \quad (5)$$

where f_{ij} indicates the frequency of allocations lying in element i,j of the 2x2 confusion matrix. The test is focused on the cases correctly classified by one classifier but mis-classified by the other. With this test, two classifications may be considered to be of different accuracy at the 95% level of confidence if $z > |1.96|$. This and related approaches has been recently advocated for use in the comparison of accuracy statements derived using the same ground data set (Foody, 2004; de Leeuw *et al.*, 2006).

5 Comparison with other mapping communities

Although the remote sensing community has gradually moved toward a position in which an accuracy assessment is seen as an essential component of a mapping exercise (Cihlar, 2000; Strahler *et al.*, 2006) this is not always the situation with other mapping communities. From the above discussion it may be apparent that sometimes the remote sensing community is rather harsh on itself by setting high standards and using techniques that commonly act to reduce the apparent accuracy of a classification. The producers of other maps, however, adopt very different approaches and criteria. Other mapping bodies, while concerned about map quality, often provide relatively little if any information on map accuracy or have relatively loosely defined and tolerant criteria of acceptability. This is not raised as a criticism of these communities or their maps as there is often good reason for the situation. However, the remote sensing community may be harsher in the evaluation of its products than other mapping communities are of theirs. The various user communities may also appear harsher in their assessments of thematic maps produced by remote sensing than other widely used maps. This is evident in relation to widely used maps.

Geological and soil maps are two widely used forms of thematic map. Both, however, may be expected to have a low accuracy if evaluated from the harsh site specific approach used in remote sensing. For example, the British Geological Survey claims that its maps are amongst the most accurate geological maps in the world (Smith, 2004). This may well be true but the maps are not accompanied by accuracy statements of the type commonly provided with thematic maps derived from a classification of remotely sensed imagery. The accuracy information generally available with a geological map relates predominantly to the spatial and cartographic components of the map rather than the thematic, geological, information contained. The accuracy statement generally provided with geological map data explicitly does not address the quality of the geological linework or data in general as much of this is a matter of interpretation as the geology is, of course, often concealed. However, as all geological units are either represented by a line or contained within a set of lines, the linework of the map is of fundamental importance yet its meaning is very uncertain. The boundaries plotted on geological maps are recognised explicitly as being no more than approximations which

indicate roughly where an actual boundary may occur. Moreover, the linework does not distinguish between the different types of boundary that may occur and the vast majority of the boundaries plotted on a geological map are inferred with many being little more than best guesses. The geological community is no doubt aware of the general nature of these maps, including their limitations, factoring this information into its work when using them. Such maps are, however, clearly likely to contain error when viewed from the overly harsh site-specific approach to thematic map accuracy assessment used in remote sensing.

As with geological mapping, there is considerable dependence on interpretation in the mapping of soils. Additionally, the classes depicted on a soil map are also often rather uncertainly defined. For example, in the UK a soil map generally shows the dominant soil series (Curtis *et al.*, 1976). Thus, a mapped polygon might be dominated by one class but some of its area may comprise a number of other soil classes. Moreover, the amount of inclusion is not always evident to the map user. Sometimes polygons may contain substantial mixtures of soil types and simply be represented in the map as mixtures. In the USDA's soil maps, up to 25% (very occasionally >50%) of a mapped polygon may actually be of a type other than that labelled (Soil Survey Division Staff, 1993). Consequently, a large proportion of the total mapped area may, therefore, be mis-labeled. Just as with geological maps, relatively little information on thematic accuracy is provided and there is considerable potential for error when viewed from the harsh site-specific perspective adopted in remote sensing. The evaluation of soil map accuracy is, however, seen as a research topic and, as recognised in other mapping communities (Maling, 1989), one that could benefit from reference to methods used in remote sensing (McBratney *et al.*, 2003).

Finally, the accuracy of topographic maps, perhaps the most widely used form of mapped information and the main alternative form of map to thematic maps, should be considered. The quality of such maps is typically evaluated in terms of a range of variables such as positional accuracy, completeness and attribute accuracy (Maling, 1989; Thapa and Bossler, 1992) but the key focus is typically on the representation of the relief and key physical features of the landscape. Accuracy statements, therefore, typically focus on the vertical and horizontal errors present in the topographic map. A standard practice would be to consider a map as being accurate if it satisfied a conventional set of map accuracy standards. For example, in relation to positional accuracy, a topographic map is normally considered to be accurate if the horizontal and vertical errors contained are below some specified threshold levels. This is a degree of tolerance not present in the approach typically used in remote sensing. Although it is difficult to compare the map evaluation approaches used for topographic and thematic mapping, it does seem that the approach used for topographic maps is less harsh than that applied to thematic maps.

Despite the problems with other maps, the remote sensing community often appears to readily use such maps unquestioningly. For example, geological, soil and topographic maps are often used in support of the production of a thematic map from remotely sensed imagery. Frequently, for example, for topographic maps to be used in pre-processing imagery, especially for geometric and topographic corrections (e.g. Hale and Rock, 2003). Geological, soil and other maps may also be used as ancillary information to help increase class separability and thereby accuracy of the classification used in the derivation of a thematic map (e.g. Loveland *et al.*, 1991; Maselli *et al.*, 1996; Bruzzone *et al.*, 1997; Homer *et al.*, 1997; Vogelmann *et al.*, 1998; Rogan *et al.*, 2003). While information on the quality of such data can sometimes be incorporated directly into a classification analysis (e.g. Peddle, 1995) such

ancillary data are commonly used directly, as if error-free, even if the analyst is aware of some possible limitations (Mas, 2004). It, therefore, seems that the remote sensing community is often prepared to accept other maps as being of acceptable quality yet is unduly harsh in the assessment of its own thematic maps.

6 Conclusions

Accuracy assessment is a fundamental component of a thematic mapping programme. The research and user communities, including the remote sensing community, often seems to be unfairly harsh in the assessment of thematic maps derived from remote sensing. This is evident in relation to the target accuracy commonly specified, the methods of accuracy assessment and comparison that are widely promoted as well as in comparison to accuracy assessment methods used by other mapping communities.

The 85% target accuracy that is often adopted in the evaluation of a thematic map derived from remotely sensed imagery appears to stem from early research on mapping broad land cover classes at a small cartographic scale and may be inappropriate for other mapping scenarios. Despite this, the 85% target value is widely used in a diverse range of thematic mapping application scenarios. In working to this target accuracy, the use of site-specific accuracy assessment methods based on the confusion matrix are also commonly used although these are often based on assumptions that are untenable (e.g. no mis-location error) and unfair (e.g. that the ground data are error-free). Furthermore, some commonly promoted measures of accuracy may unnecessarily remove chance agreement leading to an apparent reduction in map accuracy. Commonly, therefore, the evaluation of a thematic map derived from remotely sensed imagery is based on what may be considered to be an ambitiously high target accuracy of 85% and an approach to accuracy assessment that is geared to provide a pessimistically biased estimate is used. The remote sensing community may, therefore, often be chasing an unrealistic and inappropriate target and compounding the problem by using pessimistically biased techniques. From this perspective it is not surprising that many thematic maps derived from remote sensing are viewed as being of inadequate accuracy. Other types of map that are widely used without question of their accuracy may, however, also fail to satisfy a similar target if evaluated from the harsh perspective used in remote sensing. Despite this, such maps are often used without question. It, therefore, appears that the remote sensing community often appears to have a somewhat masochistic tendency in accuracy assessment, subjecting its thematic maps to an overly harsh and critical appraisal using pessimistically biased techniques yet accepting other maps with little question to their accuracy. There is, of course, a need to ensure that sources of optimistic bias in accuracy assessment (e.g. Hammond and Verbyla, 1996) are recognised in order to ensure that maps of low quality are not viewed acceptable.

There appears to be a need for a critical appraisal of some fundamental issues, such as the aims in mapping and an awareness of how realistic they are within their context. This may help to reduce unfair criticism of thematic maps that are sometimes based on false perceptions of map quality inferred from classification accuracy statements. A realistic target should be defined for each specific mapping exercise. The specification of the target value should recognise the particular features of the specific mapping task in-hand (e.g. the nature of the remotely sensed imagery used and level of class detail). This is actually similar to what Anderson *et al.* (1976) proposed for land cover mapping, in which a well-justified case for a target was specified. The main application scenario of Anderson *et al.* (1976), from which the widely used 85% target accuracy appears to have arisen, differs greatly from that of other mapping programmes.

Despite this, many map evaluation analyses have adopted the 85% target even though there is no reason to believe that the target value suggested by Anderson *et al* (1976) should be applicable to any other scenario than that they defined. Critically, there does not appear to be a universally applicable standard target to adopt and a 'one size fits all' mentality may lead to unrealistic assessments of suitability. There is also a need to recognise that problems in accuracy assessment can be a source of pessimistic bias. In particular, the rigid use of site-specific accuracy assessment methods in which all error is seen as arising from the image classification and the use of inappropriate measures to quantify and compare accuracy can lead to a misrepresentation of classification quality and mis-interpretation.

The assessment, evaluation and comparison of thematic maps are very much a topic for further research (Rindfuss *et al.*, 2004; Strahler *et al.*, 2006). There are many other issues, such as the minimum mapping unit and the unit for accuracy assessment and reporting as well as those associated with issues such as variation in error severity and the assessment of soft classifications, which require further attention.

Acknowledgements

This article is based on work undertaken over many years and has, therefore, benefited from inputs from a range of sources which is gratefully acknowledged.

References

- Abeyta, A. M. and Franklin, J., 1998. The accuracy of vegetation stand boundaries derived from image segmentation in a desert environment, *Photogrammetric Engineering and Remote Sensing*, 64, pp. 59-66.
- Agresti, A., 1996. *An Introduction to Categorical Data Analysis*, New York: Wiley.
- 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.
- Bektas, F. and Goksel, C., 2004. Remote sensing and GIS integration for land cover analysis, a case study: Gokceada island, *Proceedings XXth ISPRS Congress*, Istanbul.
- Bradley, J. V., 1968. *Distribution-free Statistical Tests*, New Jersey: Prentice-Hall.
- Brown, M., Lewis, H. G. and Gunn, S. R., 2000. Linear spectral mixture models and support vector machines for remote sensing, *IEEE Transactions on Geoscience and Remote Sensing*, 38, pp. 2346-2360.
- Bruzzone, L., Conese, C., Maselli, F. and Roli, F., 1997. Multisource classification of complex rural areas by statistical and neural-network approaches, *Photogrammetric Engineering and Remote Sensing*, 63, pp. 523-533.
- Cihlar, J., 2000. Land cover mapping of large areas from satellites: status and research priorities, *International Journal of Remote Sensing*, 21: 1093-1114.
- Cohen, J., 1960. A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, 20, pp. 37-46.
- Comber, A., Fisher, P. and Wadsworth, R., 2005. What is land cover? *Environment and Planning B*, 32, pp. 199-209.
- Congalton, R. G. and Mead, 1983. A quantitative method to test for consistency and correctness in photointerpretation, *Photogrammetric Engineering and Remote Sensing*, 49, pp. 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, pp. 1671-1678.
- Congalton, R. G. and Green, K., 1999. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, Boca Raton, Lewis Publishers.
- Curtis, L. F., Courtney, F. M. and Trudgill, S., 1976. *Soils in the British Isles*, Longman, London.

- Di Eugenio, B. and Glass, M., 2004. The kappa statistic: a second look, *Computational Linguistics*, 30, pp. 95-101.
- de Leeuw, J. Jia, H., Yang, L., Liu, X., Schmidt, K. and Skidmore, A. K., 2006. Comparing accuracy assessments to infer superiority of image classification methods, *International Journal of Remote Sensing*, 27, pp. 223-232.
- Donner, A., Shoukri, M. M., Klar, N. and Bartfay, E., 2000. Testing the equality of two dependent kappa statistics, *Statistics in Medicine*, 19, pp. 393-387.
- Fisher, P. F. and Langford, M., 1996. Modelling sensitivity to accuracy in classified imagery: a study of areal interpolations by dasymetric mapping, *Professional Geographer*, 48, 299-309.
- Fitzgerald, R. W. and Lees, B. G., 1994. Assessing the classification accuracy of multisource remote sensing data, *Remote Sensing of Environment*, 47, pp. 362-368.
- Footy, G. M., 1992. On the compensation for chance agreement in image classification accuracy assessment, *Photogrammetric Engineering and Remote Sensing*, 58, pp. 1459-1460.
- Footy, G. M., 2002. Status of land cover classification accuracy assessment, *Remote Sensing of Environment*, 80, pp. 185-201.
- Footy, G. M., 2004. Thematic map comparison: evaluating the statistical significance of differences in classification accuracy, *Photogrammetric Engineering and Remote Sensing*, 70, pp. 627-633.
- Hagen, A., 2003. Fuzzy set approach to assessing similarity of categorical maps, *International Journal of Geographical Information Science*, 17, 235-249.
- Hale, S. R. and Rock, B. N., 2003. Impact of topographic normalization on land-cover classification accuracy, *Photogrammetric Engineering and Remote Sensing*, 69, pp. 785-791.
- Hammond, T. O. and Verbla, D. L., 1996. Optimistic bias in classification accuracy assessment, *International Journal of Remote Sensing*, 17, 1261-1266.
- Hayes, D. J. and Sader, S. A., 2001. Comparison of change detection techniques for monitoring tropical forest clearing and vegetation regrowth in a time series, *Photogrammetric Engineering and Remote Sensing*, 67, pp. 1067-1075.
- Homer, C. G., Ramsey, R. D., Edwards, T. C. and Falconer, A., 1997. Landscape cover-type modeling using a multi-scene thematic mapper mosaic, *Photogrammetric Engineering and Remote Sensing*, 63, pp. 59-67.
- Janssen, L. L. F. and van der Wel, F. J. M., 1994. Accuracy assessment of satellite derived land-cover data: a review, *Photogrammetric Engineering and Remote Sensing*, 60, pp. 419-426.
- Joria, P. E. and Jorgenson, J. C. 1996. Comparison of three methods for mapping tundra with Landsat digital data, *Photogrammetric Engineering and Remote Sensing*, 62, pp. 163-169.
- Jung, H-W., 2003. Evaluating interrater agreement in SPICE-based assessments, *Computer Standards and Interfaces*, 25, pp. 477-499.
- Kerr, J. T. and Cihlar, J., 2004. Land use mapping, *Encyclopedia of Social Measurement*, Elsevier.
- Khorrarn, S. (Ed), 1999. *Accuracy Assessment of Remote Sensing-Derived Change Detection*, American Society for Photogrammetry and Remote Sensing, Bethesda MD.
- Laba, M., Gregory, S. K., Braden, J., Ogurcak, D., Hill, E., Fegraus, E., Fiore, J. and DeGloria, S. D., 2002. Conventional and fuzzy accuracy assessment of the New York Gap Analysis Project land cover map, *Remote Sensing of Environment*, 81, pp. 443-455.
- Loveland, T. R., Merchant, J. W., Ohlen, D. O. and Brown, J. F., 1991. Development of a land-cover characteristics database for the conterminous U. S., *Photogrammetric Engineering and Remote Sensing*, 57, pp. 1453-1463.
- Loveland, T. R., Zhu, Z., Ohlen, D. O., Brown, J. F., Reed, B. C. and Yang, L., 1999. An analysis of the IGBP global land-cover characterisation process, *Photogrammetric Engineering and Remote Sensing*, 65, pp. 1021-1032.
- Maling, D. H., 1989. *Measurements from Maps*, Oxford: Pergamon.
- Manel, S., Williams, C. and Ormerod, S. J., 2001. Evaluating presence-absence models in ecology: the need to account for prevalence, *Journal of Applied Ecology*, 38, pp. 921-931.

- Mas, J. F., 2004. Mapping land use/cover in a tropical coastal area using satellite sensor data, GIS and artificial neural networks, *Estuarine, Coastal and Shelf Science*, 59, pp. 219-230.
- Maselli, F., Petkov, L., Maracchi, G. and Conese, C., 1996. Eco-climatic classification of Tuscany through NOAA-AVHRR data, *International Journal of Remote Sensing*, 17, pp. 2369-2384.
- McBratney, A. B., Mendonca Santos, M. L. and Minasny, B., 2003. On digital soil mapping, *Geoderma*, 117, pp. 3-52.
- McCormick, C. M., 1999. Mapping exotic vegetation in the everglades from large-scale aerial photographs, *Photogrammetric Engineering and Remote Sensing*, 65, pp. 179-184.
- McKenzie, D. P., Mackinnon, A. J., Peladeau, N., Onghena, P., Bruce, P. C., Clarke, D. M., Haarrigan, S. and McGorry, P. D., 1996. Comparing correlated kappas by resampling: is one level of agreement significantly different from another?, *Journal of Psychiatric Research*, 30, pp. 483-492.
- Monserud, R. A. and Leemans, R., 1992. Comparing global vegetation maps with the Kappa statistic, *Ecological Modelling*, 62, pp. 275-293.
- Nishii, R. and Tanaka, S., 1999. Accuracy and inaccuracy assessments in land-cover classification, *IEEE Transactions on Geoscience and Remote Sensing*, 37, pp. 491-498.
- Peddle, D. R., 1995. Mercury \oplus : An evidential reasoning image classifier, *Computers and Geosciences*, 21, pp. 1163-1176.
- Pontius, R. G., 2000. Quantification error versus location error in comparison of categorical maps, *Photogrammetric Engineering and Remote Sensing*, 66, pp. 1011-1016.
- Pontius, R. G., 2002. Statistical methods to partition effects of quantity and location during comparison of categorical maps at multiple resolutions, *Photogrammetric Engineering and Remote Sensing*, 68, pp. 1041-1049.
- Pontius, R. G. and Cheuk, M. L., 2006. A generalised cross-tabulation matrix to compare soft-classified maps at multiple resolutions. *International Journal of Geographical Information Science*, 20, pp. 1-30.
- Rindfuss, R. R., Walsh, S. J., Turner II, B. L., Fox, J. and Mishra, V., 2004. Developing a science of land change: challenges and methodological issues, *Proceedings of the National Academy of Sciences USA*, 101, pp. 13976-13981.
- Rogan, J., Miller, J., Stow, D., Franklin, J., Levien, L. and Fischer, C., 2003. Land-cover change monitoring with classification trees using Landsat TM and ancillary data, *Photogrammetric Engineering and Remote Sensing*, 69, pp. 793-804.
- Rosenfield, G. H. and Fitzpatrick-Lins, K., 1986. A measure of agreement as a measure of thematic classification accuracy, *Photogrammetric Engineering and Remote Sensing*, 52, pp. 223-227.
- Sader, S. A., Hayes, D. J., Hepinstall, J. A., Coan, M. and Soza, C., 2001. Forest change monitoring of a remote biosphere reserve, *International Journal of Remote Sensing*, 22, pp. 1937-1950.
- Scepan, J., 1999. Thematic validation of high-resolution global land-cover data sets, *Photogrammetric Engineering and Remote Sensing*, 65, pp. 1051-1060.
- Smith, A., 2004. *Accuracy of BGS Legacy Digital Geological Map Data*, British Geological Survey, Keyworth, Nottingham.
- Smits, P. C., Dellepiane, S. G. and Schowengerdt, R. A., 1999. Quality assessment of image classification algorithms for land-cover mapping: a review and proposal for a cost-based approach, *International Journal of Remote Sensing*, 20, pp. 1461-1486.
- Soil Survey Division Staff, 1993. *Soil Survey Manual*, Soil Conservation Service, U.S. Department of Agriculture, Handbook 18.
- Stehman, S. V., 1995. Thematic map accuracy assessment from the perspective of finite population sampling, *International Journal of Remote Sensing*, 16, pp. 589-593.
- Stehman, S. V., 1997. Selecting and interpreting measures of thematic classification accuracy, *Remote Sensing of Environment*, 62, pp. 77-89.
- Stehman, S. V., Wickham, J. D., Smith, J. H. and Yang, L., 2003. Thematic accuracy of the 1992. National Land-Cover Data for the eastern United States: Statistical methodology and regional results, *Remote Sensing of Environment*, 86, pp. 500-516.
- Strahler, A. H., Boschetti, L., Foody, G. M., Friedl, M. A., Hansen, M. C., Herold, M., Mayaux, P., Morisette, J. T., Stehman, S. V. and Woodcock, C. E., 2006. *Global Land Cover Validation*:

- Recommendations for Evaluation and Accuracy Assessment of Global Land Cover Maps*, Technical Report, Joint Research Centre, Ispra, in press.
- Thapa, K. and Bossler, J., 1992. Accuracy of spatial data used in geographic information systems, *Photogrammetric Engineering and Remote Sensing*, 58, pp. 835-841.
- Thomas, I. L. and Allcock, G. McK., 1984. Determining the confidence level for a classification, *Photogrammetric Engineering and Remote Sensing*, 50, pp. 1491-1496.
- Townshend, J. R. G., 1992. Land cover, *International Journal of Remote Sensing*, 13, pp. 1319-1328.
- Treitz, P. and Rogan, J., 2004. Remote-sensing for mapping and monitoring land-cover and land-use change – an introduction, *Progress in Planning*, 61, pp. 269-279.
- Trodd, N. M., 1995. Uncertainty in land cover mapping for modelling land cover change, *Proceedings RSS95: Remote Sensing in Action*, Nottingham: Remote Sensing Society, pp. 1138-1145.
- Turk G., 2002. Map evaluation and "chance correction", *Photogrammetric Engineering and Remote Sensing*, 68, pp. 123-+
- Vogelmann, J. E., Sohl, T. and Howard, S. M., 1998. Regional characterization of land cover using multiple sources of data, *Photogrammetric Engineering and Remote Sensing*, 64, pp. 45-57.
- Weng, Q., 2002. Land use change analysis in the Zhujiang delta of China using satellite remote sensing, GIS and stochastic modelling, *Journal of Environmental Management*, 64, pp. 273-284.
- Wheeler, A. P. and Allen, M. S., 2002. Comparison of three statistical procedures for classifying the presence-absence of an aquatic macrophyte from microhabitat observations, *Journal of Freshwater Ecology*, 17, pp. 601-608.
- Wilkinson, G. G., 1996. Classification algorithms - where next? *Soft Computing in Remote Sensing Data Analysis*, (E. Binaghi, P. A. Brivio and A. Rampini, editors), World Scientific, Singapore, pp. 93-99.
- Wilkinson, G. G., 2005. Results and implications of a study of fifteen years of satellite image classification experiments, *IEEE Transactions on Geoscience and Remote Sensing*, 43, pp. 433-440.
- Wright, G. G. and Morrice, J. G., 1997. Landsat TM spectral information to enhance the land cover of Scotland 1988 data set, *International Journal of Remote Sensing*, 18, pp. 3811-3834.
- Wulder, M. A., Franklin, S. E., White, J. C., Linke, J. and Magnussen, S., 2006. An accuracy assessment framework for large-area land cover classification products derived from medium-resolution satellite data, *International Journal of Remote Sensing*, 27, pp. 663-683.