ELSEVIER

Remote Sensing of Environment 80 (2002) 185–201

Remote Sensing of Environment

www.elsevier.com/locate/rse

Status of land cover classification accuracy assessment

Giles M. Foody*

Department of Geography, University of Southampton, Highfield, Southampton, SO17 1BJ, UK

Received 12 March 2001; received in revised form 13 July 2001; accepted 21 July 2001

Abstract

The production of thematic maps, such as those depicting land cover, using an image classification is one of the most common applications of remote sensing. Considerable research has been directed at the various components of the mapping process, including the assessment of accuracy. This paper briefly reviews the background and methods of classification accuracy assessment that are commonly used and recommended in the research literature. It is, however, evident that the research community does not universally adopt the approaches that are often recommended to it, perhaps a reflection of the problems associated with accuracy assessment, and typically fails to achieve the accuracy targets commonly specified. The community often tends to use, unquestioningly, techniques based on the confusion matrix for which the correct application and interpretation requires the satisfaction of often untenable assumptions (e.g., perfect coregistration of data sets) and the provision of rarely conveyed information (e.g., sampling design for ground data acquisition). Eight broad problem areas that currently limit the ability to appropriately assess, document, and use the accuracy of thematic maps derived from remote sensing are explored. The implications of these problems are that it is unlikely that a single standardized method of accuracy assessment and reporting can be identified, but some possible directions for future research that may facilitate accuracy assessment are highlighted.

1. Introduction

Land cover is a fundamental variable that impacts on and links many parts of the human and physical environments. Land cover change is, for example, regarded as the single most important variable of global change affecting ecological systems (Vitousek, 1994) with an impact on the environment that is at least as large as that associated with climate change (Skole, 1994). It is well established that land cover change has significant effects on basic processes including biogeochemical cycling and thereby on global warming (Penner, 1994), the erosion of soils and thereby on sustainable land use (Douglas, 1999), and for at least the next 100 years is likely to be the most significant variable impacting on biodiversity (Chapin et al., 2000). Despite the significance of land cover as an environmental variable, our knowledge of land cover and its dynamics is poor. Understanding the significance of land cover and predicting the effects of land cover change is particularly limited by the paucity of accurate land cover data. Such data, especially in map form, are, contrary to popular belief in some quarters, not readily available or trivially easy to acquire (DeFries & Townshend, 1994; Estes & Mooneyhan, 1994; Rhind & Hudson, 1980). Moreover, few items depicted on maps change as rapidly as land cover and so an ability to monitor it accurately is important (Belward, Estes, & Kilne, 1999).

Remote sensing is an attractive source of thematic maps such as those depicting land cover as it provides a map-like representation of the Earth's surface that is spatially continuous and highly consistent, as well as available at a range of spatial and temporal scales. Thematic mapping from remotely sensed data is typically based on an image classification. This may be achieved by either visual or computer-aided analysis. The classification may be one that seeks to group together cases by their relative spectral similarity (unsupervised) or that aims to allocate cases on the basis of their similarity to a set of predefined classes that have been characterized spectrally (supervised). In each situation, the resulting classified image may be treated as a thematic map depicting the land cover of the region. Although remote sensing has been used successfully in mapping a range of land covers at a variety of spatial and temporal scales, its full potential as a source of land cover information has not been realized (Townshend, 1992; Wil-

* Fax: +44-1703-593-295.

E-mail address: g.m.foody@soton.ac.uk (G.M. Foody).

0034-4257/01/$ – see front matter
PII: S0034-4257(01)00295-4

kinson, 1996). A key concern is that the land cover maps derived are often judged to be of insufficient quality for operational applications. This judgement is typically based on an evaluation of the derived land cover map against some ground or other reference data set. Disagreements between the two data sets are typically interpreted as errors in the land cover map derived from the remotely sensed data (Congalton, 1991; Smedes, 1975). This interpretation has driven research that aims to decrease the error in image classification. This research has typically focused on the derivation and assessment of different classification algorithms. It has also led to the questioning of the spectral and radiometric suitability of remotely sensed data sets used in thematic mapping applications (Estes et al., 1999; Wilkinson, 1996) and the use of classification methods as the tool in mapping from remotely sensed data (Foody, 1999; Mather, 1999). However, there are many uncertainties associated with the meaning and interpretation of map quality that make it a difficult variable to consider objectively and which substantially limit the ability to evaluate the degree to which the potential of remote sensing as a source of land cover data is being realized. As with a range of geospatial data sets, there is generally a lack of information on data quality and what there is may be poorly communicated to the user (Johnston & Timlin, 2000).

The quality of spatial data sets is a very broad issue that may relate to a variety of properties (Worboys, 1998) but frequently, and here, the property of interest is map or classification accuracy. As any map is simply a model or generalization, it will contain error (Brown, Loveland, Ohlen, & Zhu, 1999; Dicks & Lo, 1990; Maling, 1989; Smits, Dellepiane, & Schowengerdt, 1999). Thus, although a thematic map provides a typically unquestioned simplification of reality, it has flaws and is only one model or representation of the depicted theme (Woodcock & Gopal, 2000). For example, as the mapping processes involves generalization there is some loss of information and so completeness (Maling, 1989). It is important, therefore, that the quality of thematic maps derived from remotely sensed data be assessed and expressed in a meaningful way. This is important not only in providing a guide to the quality of a map and its fitness for a particular purpose, but also in understanding error and its likely implications, especially if allowed to propagate through analyses linking the map to other data sets (Arbia, Griffith, & Haining, 1998; Janssen & van der Wel, 1994; Veregin, 1994). Although classification accuracy assessment is now widely accepted as a fundamental component of thematic mapping investigations (Cihlar, 2000; Cohen & Justice, 1999; Congalton, 1994; Justice et al., 2000; Merchant, Yang, & Yang, 1994), it is not uncommon for map accuracy to be inadequately quantified and documented (Dicks & Lo, 1990). There may be many reasons for this situation. Although it may appear simple in concept, accuracy is a difficult property to measure and express. This paper aims to briefly review the status of land cover classification accuracy assessment and point the interested reader to some of the extensive literature on the topic. It will consider how accuracy is assessed and expressed before looking at some of the problems encountered and suggestions of future research priorities.

2. Background to classification accuracy assessment

In a statistical context, accuracy comprises bias and precision and the distinction between the two is sometimes important as one may be traded for the other (Campbell, 1996; Maling, 1989). In thematic mapping from remotely sensed data, the term accuracy is used typically to express the degree of 'correctness' of a map or classification. A thematic map derived with a classification may be considered accurate if it provides an unbiased representation of the land cover of the region it portrays. In essence, therefore, classification accuracy is typically taken to mean the degree to which the derived image classification agrees with reality or conforms to the 'truth' (Campbell, 1996; Janssen & van der Wel, 1994; Maling, 1989; Smits et al., 1999). A classification error is, thus, some discrepancy between the situation depicted on the thematic map and reality.

In early mapping studies, accuracy assessment was frequently an afterthought (Congalton & Green, 1993; Jensen, 1996). This may have been exacerbated by the relative ease of producing accurate looking maps from digital image processors (Dicks & Lo, 1990). Accuracy assessment has, however, evolved considerably. The history of accuracy assessment reveals increasing detail and rigor in the analysis. Congalton (1994) identifies four major historical stages in accuracy assessment. In the first, accuracy assessment was based on a basic visual appraisal of the derived map. The map would be considered accurate, essentially, if it looked right or good. Clearly, such a highly subjective approach is often inappropriate, particularly given the ability to derive polished outputs from digital image processing systems. The second historical stage was characterized by an attempt to quantify accuracy more objectively. In this, accuracy assessment was based on comparisons of the areal extent of the classes in the derived thematic map (e.g., km^2 or % cover of the region mapped) relative to their extent in some ground or other reference data set. The nonsite-specific nature of this approach is, however, a major limitation as a map could easily display the classes in the correct proportions but in the incorrect locations, greatly limiting the value of the map for some users. Thus, a major problem with this approach to accuracy assessment is that the apparent and quantified accuracy of the map would hide its real quality. The third stage in the history of accuracy assessment involved the derivation of accuracy metrics that were based on a comparison of the class labels in the thematic map and ground data for a set of specific locations. These site-specific approaches include measures such as the percentage of cases correctly allocated (sometimes referred to as overall accuracy). Finally, the

fourth stage in the history of accuracy assessment is a refinement of the third in which greater use of the information on the correspondence of the predicted thematic map labels to those observed on the ground is made. This stage has the confusion or error matrix at its core and uses this to describe the pattern of class allocation made relative to the reference data. A key characteristic feature of this stage is that measures of accuracy that use the information content of the confusion matrix more fully than the basic percentage of correctly allocated cases, such as the kappa coefficient of agreement, are frequently derived to express classification accuracy. Presently, the confusion or error matrix is at the core of accuracy assessment but there is much scope to extend the analysis beyond it (Congalton, 1994; Congalton & Green, 1999).

The history of accuracy assessment outlined above, however, relates mainly to mapping investigations that have focused on local to regional scales. The methods used may not be transferable to coarser scales (Merchant et al., 1994). In recent years, however, there has been increasing demand for remote sensing to provide maps at all scales but especially at regional to global scales. As studies have increasingly focused on mapping very large areas, the nature of the mapping problem has in some ways changed. Necessarily, regional to global scale mapping is often constrained to use relatively coarse spatial resolution data, such as from the NOAA AVHRR, due to constraints of data cost, volume, and the relatively high probability of obtaining a cloud-free view of the land surface with such data. Thus, major recent mapping investigations at the global scale have resulted in the production of maps with a spatial resolution of 1 km or coarser (Belward et al., 1999; DeFries, Hansen, Townshend, & Sohlberg, 1998; Hansen, DeFries, Townshend, & Sohlberg, 2000; Loveland et al., 2000; Muchoney et al., 2000). From an accuracy assessment standpoint, however, there are several major concerns arising from this situation. A key problem is a lack of accuracy standards for such maps (Loveland et al., 1999). Accuracy assessment of large area maps remains, therefore, a challenging issue (Justice et al., 2000; Stehman, Wickham, Yang, & Smith, 2000). Consequently, it is common to find that maps of large areas are often accompanied by accuracy statements that are known to be erroneous or perhaps are relatively vague or nonsite-specific. This is typically not a criticism of the map producers but evidence of major problems in classification accuracy assessment. Brutally persevering with a standard confusion matrix-based approach to accuracy assessment could, for example, be misleading as it would be unlikely that the assumptions underlying such approaches (e.g., pure pixels, discrete classes, etc.) are satisfied.

In addition to the problems associated with the mapping of large areas, there are a variety of other scale-independent issues to consider in the evaluation of the accuracy of a classification. Indeed, the measurement and meaning of classification accuracy depend considerably on one's particular viewpoint and requirements (Buckton, O'Mongain, & Danaher, 1999; Campbell, 1996; Czaplewski, 1992; Stehman, 1999a). An accuracy assessment may be undertaken for different reasons. It could, for instance, be undertaken to provide an overall measure of the quality of a map, to form the basis of an evaluation of different classification algorithms or to attempt to help gain an understanding of errors (Congalton et al., 1998; Hay, 1979; Richards, 1996). In each instance, different users may have different concerns about accuracy. They may, for example, be interested in the overall or global accuracy, the accuracy with which a specific class has been mapped, or the accuracy of area estimates. Furthermore, while some errors are of no concern to some users, they may be detrimental to others (DeFries & Los, 1999), yet rarely are misclassifications treated unequally (Naesset, 1996). Similarly, no one classification will be optimal from the viewpoint of each different user (Brown et al., 1999; Lark, 1995). As classification accuracy has various components and users differ in their specific needs, it is important to measure the desired properties (Lark, 1995). Thus, it is vital that the component of accuracy measured is appropriate for the requirements of each particular study to avoid misinterpretation (Stehman, 1997a). Consequently, seeking to optimize accuracy expressed by one metric may lead to a suboptimal classification when viewed from another standpoint and accuracy quantified with a different metric (Morisette & Khorram, 2000). These and other issues complicate the assessment and reporting of classification accuracy, which in turn limits the value of remote sensing as a source of land cover data. Before considering some of the problems in more detail, the popular basis to accuracy assessment will be briefly reviewed.

3. Promoted accuracy measures and calls for standardization

Many methods of accuracy assessment have been discussed in the remote sensing literature (e.g., Aronoff, 1982, 1985; Kalkhan, Reich, & Czaplewski, 1995; Koukoulas & Blackburn, 2001; Piper, 1983; Rosenfield & Fitzpatrick-Lins, 1986). The most widely promoted and used, however, may be derived from a confusion or error matrix.

The confusion matrix is currently at the core of the accuracy assessment literature. As a simple cross-tabulation of the mapped class label against that observed in the ground or reference data for a sample of cases at specified locations, it provides an obvious foundation for accuracy assessment (Campbell, 1996; Canters, 1997). Indeed, the confusion matrix provides the basis on which to both describe classification accuracy and characterize errors, which may help refine the classification or estimates derived from it. For example, the matrix may reveal interclass confusion that could be resolved with the use of additional discriminatory information. Alternatively, the pattern of misclassification evident in the matrix may aid studies that use the map, particularly as a means to estimating the areal

extent of classes over a region (Czaplewski, 1992; Hay, 1988; Jupp, 1989; Prisley & Smith, 1987; Van Deusen, 1996; Yuan, 1997).

Many measures of classification accuracy may be derived from a confusion matrix. One of the most popular is the percentage of cases correctly allocated. This is an easily interpretable guide to the overall accuracy of the classification. If attention focuses on the accuracy of individual classes, then the percentage of cases correctly allocated may be derived from the confusion matrix by relating the number of cases correctly allocated to the class to the total number of cases of that class. This may be achieved from two standpoints, giving rise to the somewhat awkwardly termed (Janssen & van der Wel, 1994) user's and producer's accuracy, depending on whether the calculations are based upon the matrix's row or column marginals (Campbell, 1996; Story & Congalton, 1986). The calculation of these, and some other major indices, is illustrated in Fig. 1 for data obtained via simple random sampling. In the derivation of these indices, a number of fundamental assumptions are typically made. For example, it is generally assumed implicitly that each case (e.g., pixel) to be classified belongs fully to one of the classes in an exhaustively defined set of discrete and mutually exclusive classes (Congalton et al., 1998; Congalton & Green, 1999; Lewis & Brown, in press; Townsend, 2000).

Although informative, measures such as the percentage of cases correctly classified have often been criticized. A major problem for some users is that some cases may have been allocated to the correct class purely by chance (Congalton, 1991; Pontius, 2000; Rosenfield & Fitzpatrick-Lins, 1986; Turk, 1979). To accommodate for the effects of chance agreement, Cohen's kappa coefficient has often been used and some commentators argue that it should, in some circumstances, be adopted as a standard measure of classification accuracy (Smits et al., 1999). The kappa coefficient has many attractive features as an index of classification accuracy. In particular, it makes some compensation for chance agreement and a variance term may be calculated for it enabling the statistical testing of the significance of the difference between two coefficients (Rosenfield & Fitzpatrick-Lins, 1986). This is often important, as frequently, there is a desire to compare different classifications and so matrices. To further aid this comparison, some have called for the normalization of the confusion matrix such that each row and column sums to unity (Congalton, 1991; Smits et al., 1999).

Accuracy assessment has been a topic of considerable debate and research in remote sensing for many years. This is in part because the promoted standard methods such as the kappa coefficient are not always appropriate. Moreover, there is nothing unique about the kappa coefficient in compensating for chance agreement or in allowing the significance of differences in accuracy to be evaluated as these are features shared with other accuracy metrics. As a topic, however, accuracy assessment has matured to such an extent that many have called for a standardization of both the method of assessment and style of reporting, often with target accuracy thresholds specified (e.g., Congalton, Green, & Teply, 1993; Smits et al., 1999; Thomlinson, Bolstad, & Cohen, 1999). These target accuracies often tend to be based upon the influential work of Anderson, Hardy, Roach, and Witmer (1976). Typically, the specified requirements take the form of a minimum level of overall accuracy, expressed numerically by some index such as the percentage of cases correctly allocated, and a desire for each class to be classified to comparable accuracy. Thus, for example, Thomlinson et al. (1999) set a target of an overall accuracy of 85% with no class less than 70% accurate. Additional features typically called for are the provision of more than one measure of classification accuracy (Muller et al., 1998; Stehman, 1997a), with associated confidence limits (Stehman, 1997a; Thomas & Allcock, 1984), together with the confusion matrix (Stehman, 1997a), sometimes normalized (Congalton, 1991; Smits et al., 1999). Thus, although there is no set standard method, there is a fair degree of consensus about the general format that accuracy assessment and reporting should take, with some recommended techniques seen as virtual standards and are widely adopted (Congalton, 1994). Indeed, some regard the remote sensing community as being ripe for further standardization in regard to accuracy assessment and reporting (Smits et al., 1999).

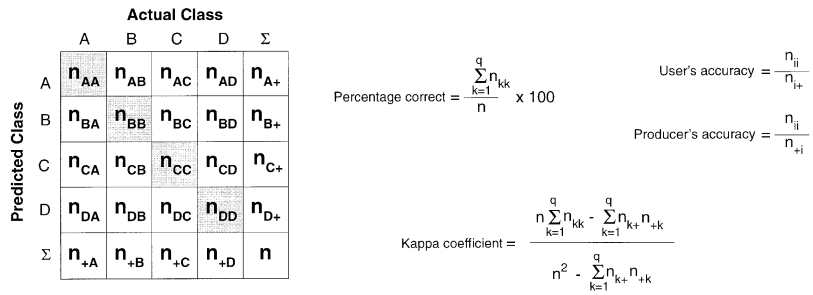

Fig. 1. The confusion matrix and some common measures of classification accuracy that may be derived from it. The highlighted elements represent the main diagonal of the matrix that contains the cases where the class labels depicted in the image classification and ground data set agree, whereas the off-diagonal elements contain those cases where there is a disagreement in the labels. In the example shown, the number of classes, q, is 5. Note that simple random sampling has been assumed; alternative formulae exist for matrices based on different sampling designs.

4. Accuracy assessment in reality

Despite the attractions of the recommended standard methods of accuracy assessment and reporting, it seems that the remote sensing community has not heeded the calls to adopt them and often does not achieve the typically specified targets. The failure to attain the specified target levels of accuracy is typically taken to indicate a failure of remote sensing as a source of land cover information (Smedes, 1975). There are, however, many problems with the assessment and reporting of classification accuracy that complicate the interpretation of accuracy statements. Some of these problems will be discussed in Section 5 after first making some observations on the methods of accuracy assessment used and the levels of accuracy achieved by researchers.

Trodd (1995) reviewed the methods used in evaluating image classifications in a survey of 84 classifications reported in 25 papers published in major journals in 1994–1995, comfortably within the fourth stage of accuracy assessment (the 'age of the error matrix') outlined by Congalton (1994). Trodd found that only 60% of the papers provided a confusion matrix and only 44% gave two or more quantitative measures of accuracy. The percentage of cases correctly allocated (the central measure of the third historical stage in accuracy assessment) was the most widely used measure of accuracy, used in 68% of papers, while the kappa coefficient was used in only 48% of papers. Perhaps more importantly, 8% of the articles provided no quantitative measure of accuracy. In those articles in which accuracy was quantified, the accuracy of the classifications reported was also generally below the commonly recommended 85% target. In addition, it was apparent that the aim of attaining a roughly equal level of accuracy across all mapped classes was rarely obtained. Trodd, for example, found that the mean range in the producer's accuracy of the classifications reviewed was 59%. Even a cursory review of more recent papers will confirm that the trends reported by Trodd are relevant to more contemporary work. The recently released IGBP global land cover map has, for example, an area weighted accuracy of 66.9%, which is comfortably less than the specified target of 85% and the accuracy with which the individual classes are mapped ranges from 40% to 100% (Scepan, 1999). Many other studies discuss classifications with overall accuracies below the general target of 85% and have a large range in the accuracy with which the individual classes have been classified (e.g., DeGloria et al., 2000; Ung, Lambert, Guidon, & Fournier, 2000). It is also common to find that nonsite-specific accuracy measures, associated with the second stage in the history of accuracy assessment, are still used (e.g., CCRS, 1999). Indeed the 'looks good' approach that characterized early (first stage) accuracy assessments may still sometimes be considered useful (Merchant et al., 1994).

The key concern is that most of the papers reviewed by Trodd (1995), and probably in the literature as a whole, have not adopted the methods of accuracy assessment and reporting typically recommended and also generally fail to meet the typically suggested target levels of accuracy. Thus, the suggested components of the accuracy statement (e.g., a confusion matrix and a metric such as the kappa coefficient) are often not provided and the target accuracies rarely attained (e.g., relatively even levels of accuracy for all classes and the overall accuracy $>85\%$). Why has the research community not adopted the methods and approaches to accuracy assessment and reporting that are widely promoted? There are a variety of explanations for this situation ranging from ignorance through laziness to problems that arise if the recommended approaches are used. The problems encountered are the key concern here as they highlight major limitations in the mapping and monitoring of land cover from remotely sensed data as well as having important implications for the users of land cover data derived from remote sensing.

5. Problems in accuracy assessment

Although the basic approaches to accuracy assessment seem relatively straightforward, many problems are often encountered when evaluating an image classification. These problems range from issues associated with a failure to satisfy basic underpinning assumptions through to the limited amount of information on map quality that is actually conveyed by a basic accuracy assessment. Several somewhat interrelated problems that limit the quantification of classification accuracy and thereby the use of land cover maps derived from remote sensing are briefly elaborated on below.

5.1. Accuracy measure and reporting

There are many measures of accuracy that can be derived from a confusion matrix (Lark, 1995; Stehman, 1997a). Although many commentators have recommended that measures such as the kappa coefficient of agreement be adopted as a standard (e.g., Smits et al., 1999), these are not without problems (Foody, 1992; Pontius, 2000). Some have argued, for example, that chance agreement is effectively overestimated in the calculation of the kappa coefficient resulting in an underestimation of classification accuracy (Foody, 1992; Ma & Redmond, 1995) or that, as a non-probability-based measure, the kappa coefficient is an inappropriate basis for accuracy assessment (Stehman & Czaplewski, 1998). Perhaps, more fundamentally, it must be realized that the various measures of accuracy evaluate different components of accuracy and make different assumptions of the data (Lark, 1995; Stehman, 1999a). For each component of accuracy there is a set of accuracy measures that may be calculated to express it. There is, therefore, no single universally acceptable measure of accuracy but instead a variety of indices, each sensitive to different features (Stehman, 1997a). Unfortunately, the

selection of an appropriate measure is not aided by the inconsistent use of terminology in the literature and the often unclear statement of study objectives (Canters, 1997; Janssen & van der Wel, 1994).

Thus, while there are calls to standardize parts of the accuracy assessment programme, including the methods used and style of reporting (Smits et al., 1999; Thomlinson et al., 1999), it is important to recognize the variety of different needs and interpretations that exist. In reality, it is probably impossible to specify a single, all-purpose measure of classification accuracy. This is unfortunate as the desire for a single index of accuracy is large. This is not a new or unusual problem. Similar problems are encountered in other disciplines with, for example, the quantification of the commonly discussed and important variable of biodiversity torn between measures of species number (richness) and abundance (evenness). One measure alone does not provide a satisfactory description of biodiversity but provided together, perhaps combined into some dimension of biodiversity, they provide a fuller description of the variable (Purvis & Hector, 2000). This may explain why some commentators have recommended that rather than simply providing a basic accuracy statement to accompany an image classification (e.g., a statement of the overall percentage of cases correctly allocated), it may be preferable to derive more than one measure of accuracy and provide the confusion matrix as a fuller description of classification accuracy (Arbia et al., 1998; Muller et al., 1998; Stehman, 1997a). This is important as the use of different accuracy measures may result in different, possibly conflicting, interpretations and conclusions (Stehman, 1997a). It is also important that the raw confusion matrix be presented and that it must not be a normalized one as it may be inappropriate for some users. Normalization can, for example, lead to bias and would have the effect of equalizing the user's and producer's accuracies that may actually differ significantly (Stehman & Czaplewski, 1998). The user who might benefit from normalization of the matrix (e.g., Smits et al., 1999) is free to undertake this analysis while it is, of course, impossible to work the other way and derive the original matrix from the normalized version. Although the provision of a confusion matrix and two or more measures of accuracy may seem reasonable and feasible, it is, as noted above, common for accuracy to not be reported in such a way (Trodd, 1995). Even if it was, the users of the map may ignore the matrix (Fisher, 1994) or not understand it without training in its interpretation.

5.2. Sampling issues

The design of an accuracy assessment programme has several elements including the definition of an appropriate sample size and sampling design as well as the specification and use of a measure of accuracy appropriate to the application in-hand (Dicks & Lo, 1990; Stehman, 1999b). These are not trivial tasks. The sample size, for example, must be selected with care and be sufficient to provide a representative and meaningful basis for accuracy assessment (Hay, 1979). An appropriately defined sample will aid the ability to infer the properties of the population from which it was drawn. Typically, a design-based inferential framework is adopted in accuracy assessment programmes (Stehman, 2000). With such a framework, an appropriately sized sample may be estimated from conventional statistics (Cochran, 1977). Alternatively, a model-based inferential framework has sometimes been adopted (Stehman, 2000). With this basis, geostatistics may be used to design an efficient sample (Atkinson, 1991).

The sampling design used to select the cases upon which the accuracy assessment is based is of major importance. If, for example, a probability-based measure of classification accuracy is to be used (Stehman, 1997a), it is essential that the cases were acquired according to an appropriate sampling design (Hay, 1979; Stehman, 1999b). How often this is achieved is open to question as typically little information on the sampling design used in evaluating classification accuracy is provided with the accuracy statement. It is nonetheless important that the sampling design used is specified as it can significantly influence the results of an analysis (Friedl et al., 2000; Green, Strawderman, & Airola, 1993; Stehman, 1995). Indeed, the confusion matrix cannot be properly interpreted without knowledge of the sampling design used in its construction (Maling, 1989; Stehman, 1995).

Basic sampling designs, such as simple random sampling, can be appropriate if the sample size is large enough to ensure that all classes are adequately represented. The adoption of a simple sampling design is also valuable in helping to meet the requirements of a broad range of users (Stehman & Czaplewski, 1998) although the objectives of all users cannot be anticipated (Stehman et al., 2000). Often, however, it is impractical to follow such sampling procedures. For example, it may be extremely difficult to use randomly located sites to assess the accuracy of a map covering a very large area. Frequently, ground data collection is constrained as physical access to some sites is impractical and restricted to sites of opportunity, where it is possible to obtain ground data, or high-quality fine spatial resolution imagery acquired at an appropriate date is available as a surrogate for ground observation (Edwards, Moisen, & Cutler, 1998; Estes et al., 1999). Alternative sample designs may, therefore, be required and these can, under certain circumstances, be successfully employed with some popular measures of accuracy (e.g., Stehman, 1996a).

Commonly, budget or other practical constraints influence the selection of a sampling design. The strategies that have been adopted range from 'windshield' surveys to techniques based on double sampling (Kalkhan, Reich, & Stohlgren, 1998; Stehman, 1996b) and cluster sampling (Stehman, 1997b, 1999b; Todd, Gehring, & Haman, 1980). However, it must be realized that the sampling design used to collect the sample of cases, upon which the accuracy assessment is based, has important implications to

the estimation of classification accuracy. While there is an obvious desire to balance statistical requirements with practicalities (Belward et al., 1999; Congalton, 1991; Edwards et al., 1998; Merchant et al., 1994), the choice of sampling design influences the reliability of an accuracy assessment (Muller et al., 1998; Stehman, 1999b, 2001). In essence, a study's practical constraints should be considered in the design of the accuracy assessment programme. In that way, practical issues should not reduce the credibility of the accuracy statement derived (Stehman & Czaplewski, 1998).

5.3. Type of error

A variety of errors are encountered in an image classification. Typically, interest focuses on thematic accuracy, which is the correspondence between the class label assigned by the classification and that observed in reality. The confusion matrix appears to provide an excellent summary of the two types of thematic error that can occur, namely, omission and commission. Unfortunately, however, other sources of error contribute to the pattern of misclassification depicted in the confusion matrix (Canters, 1997; Congalton & Green, 1993; Husak, Hadley, & McGwire, 1999; Merchant et al., 1994; Muchoney, Strahler, Hodges, & LoCastro, 1999). Nonthematic errors may result in misrepresentation, typically underestimation, of the actual accuracy (Congalton & Green, 1993). Unfortunately, nonthematic errors can be large and particular concern focuses on errors due to misregistration of the image classification with the ground data (Canters, 1997; Czaplewski, 1992; Muller et al., 1998; Stehman, 1997a; Todd et al., 1980). If the two data sets are not accurately coregistered, the assessment of thematic accuracy is hampered and this problem is most apparent in heterogeneous landscapes with a complex land cover mosaic (Loveland et al., 1999; Scepan, 1999). Such landscapes may, however, be the focus of particular concern and where it is perhaps hardest but possibly most important to map and monitor land cover. In studies of land cover change based on a temporal sequence of remotely sensed imagery, misregistration errors can act to significantly exaggerate or alternatively mask change and so substantially limit the value of remote sensing for monitoring land cover dynamics (Roy, 2000; Schlagel & Newton, 1996).

Issues ranging from the properties of the sensor and ground to the methods used to preprocess the data can have a marked impact on the ability to accurately locate a pixel (Bastin, Edwards, & Fischer, 2000). This positional uncertainty can have a major detrimental effect on thematic mapping studies. Significant misregistration problems have often been observed in the mapping of large areas where the problems of obtaining a high positional accuracy have been noted to be a major source of classification error (Muller et al., 1998) that is sometimes larger than the actual thematic error (Riemann, Hoppus, & Lister, 2000). Misregistration of the data sets can, therefore, be a significant complication to accuracy assessment programmes. For example, in the IGBP global land cover map and the land cover map of Great Britain, approximately 20% and 48% of the testing sites used in each, respectively, were so incorrectly located that spatial adjustments were required for the assessment of classification accuracy (Fuller, Groom, & Jones, 1994; Husak et al., 1999). Such adjustments are, however, rarely reported in the literature although providing a simple ad hoc solution to a major problem. Only rarely is the horizontal positional accuracy of a map discussed as an issue (e.g., Thomlinson et al., 1999; Yang, Stehman, Wickham, Jonathan, & VanDriel, 2000). It seems more commonplace, and aided by some image processing software, for sites to be compared directly as if perfectly coregistered. Without perfect registration, however, the confusion matrix may contain errors due to misregistration as well as thematic mislabeling which will complicate the interpretation of accuracy metrics derived from it. It does, therefore, seem inappropriate for the remote sensing community to try to rigidly adopt site-specific accuracy assessment procedures when the ability to colocate sites in the derived map and ground data set is so difficult. The strict insistence upon perfect positional accuracy can only act to compound error (Maling, 1989).

Locational accuracy is a major concern in any mapping exercise yet producers of other types of maps allow a certain degree of tolerance. In topographic maps, for example, it is important that the features plotted are located correctly but limited horizontal and vertical deviation from reality is accepted. The US National Map Accuracy Standards for maps at a scale of 1:20,000 or smaller, for example, demands a horizontal accuracy such that 90% of a sample of well-defined points tested are accurate to within 1/50th of an inch on the map. At 1:24,000 scale this distance equates to an allowable error of up to 40 ft (12.2 m) for the selected sites and any magnitude of error for the rest. The vertical accuracy requirement for the map is for 90% of sites to be within one half of the contour interval represented on the map (Maling, 1989). Why cannot some level of positional tolerance be more generally incorporated into thematic map accuracy assessment?

5.4. Accuracy of the ground or reference data

Thus far, and in most of the literature, it has been assumed that the ground or reference data used in the assessment of classification accuracy are themselves an accurate representation of reality. In fact, the ground data are just another classification which may contain error (Congalton & Green, 1999; Khorram, 1999; Lunetta, Iiames, Knight, Congalton, & Mace, 2001; Zhou, Robson, & Pilesjo, 1998). These may be thematic errors in which the class labels are erroneous but may also include other errors such as those due to mislocation (Dicks & Lo, 1990). The certainty with which a label can be attached to a case is, for example, variable, often based on highly subjective interpretations (Thierry & Lowell, 2001). Variation in the con-

fidence of the class labeling in the ground data can significantly influence the apparent accuracy of a classification (Zhu, Yang, Stehman, & Czaplewski, 2000). This is evident in the IGBP global land cover map where the degree of consensus in ground data labeling could account for a 10.7% (77.6 – 66.9%) difference in the area weighted classification accuracy (Estes et al., 1999; Scepan, 1999; Scepan, Menz, & Hansen, 1999). If the ground data labels vary in confidence, it may sometimes be appropriate to apply an index of confidence to the reference data so that different subsets can be evaluated, or to use secondary class labels to allow a softer evaluation of the degree of agreement between the data sets to be calculated (Zhang & Foody, 1998; Zhu et al., 2000). Due to these and other problems, the common usage of the term 'truth' when describing ground data is problematic and should be avoided (Bird, Taylor, & Brewer, 2000; Khorram, 1999). More importantly, the accuracy assessments outlined above are actually only measuring the degree of agreement or correspondence to the ground data and so are not necessarily a true reflection of the closeness to reality (Congalton & Green, 1993; Merchant et al., 1994). A meaningful accuracy assessment clearly requires that the ground data are accurate. However, ground data sets often contain error and sometimes possibly more error than the remotely sensed product they are being used to evaluate (Abrams, Bianchi, & Pieri, 1996; Bauer et al., 1994; Bowers & Rowan, 1996; Merchant et al., 1994). In the interpretation of classification accuracy, however, disagreement between the derived land cover map and the ground data is typically, and unfairly, taken to indicate an error in the map derived from the remotely sensed data when other explanations exist (Congalton, 1991; Fitzgerald & Lees, 1994; Smedes, 1975). As long as the ground data are assumed to be more accurate than the derived thematic map, there is a danger of allowing a self-fulfilling prophecy to develop in which all error is associated with the thematic map simply because it was assumed to be the least accurate data set at the outset (Khorram, 1999).

Problems with ground data accuracy may be particularly severe if a remotely sensed data set is used as the reference data. Unfortunately, the use of remotely sensed data as reference data is common in the 'validation' of coarse spatial resolution map products depicting very large areas (e.g., Justice et al., 2000; Thomlinson et al., 1999). In such situations, the absence of actual ground-observed ground data and the practical constraints of the scale of the study forces the accuracy assessment to be based upon a comparison of the derived map with one derived from an image with a finer spatial resolution. It must, however, be recognized that the resulting confusion matrix and accuracy statement may be significantly distorted by errors in the reference data. In some studies it may, therefore, be important to know the methods and protocols (including issues such as class definitions) of reference data acquisition as this may influence their accuracy and suitability for relation to the thematic map in order to assess its accuracy (Bird et al., 2000; Czaplewski, 1992; Scepan et al., 1999; Zhou et al., 1998).

A further problem arises as a consequence of the sampling strategy adopted in some ground/reference data collection programmes. The size or support of the sampling units used in ground data collection is often different to the units mapped from the imagery (e.g., pixels or parcels) leading to difficulties in analyzing the data sets (Atkinson, Foody, Curran, & Boyd, 2000). The comparison of ground and thematic map labels may, therefore, be based upon differently sized units, which can result in different estimates of classification accuracy (Biging, Colby, & Congalton, 1999). Furthermore, accuracy may be assessed using a range of spatial units and the unit selected can have a major impact on the estimated magnitude of classification accuracy (Zhu et al., 2000). For example, a 23.1% difference in accuracy was reported by Yang et al. (2000) arising as a consequence of the definition of agreement between the map and ground data labels that was adopted. Of particular importance are the size and purity of the minimum mapping unit used and how it is related to the nature of the ground data (Biging et al., 1999; Khorram, 1999).

Finally, sampling is often consciously constrained to large homogeneous regions of the classes with regions in and around the vicinity of complexities such as boundaries excluded (Dicks & Lo, 1990; Mickelson, Civco, & Silander, 1998; Richards, 1996; Wickham, O'Neill, Ritters, Wade, & Jones, 1997), frequently as a deliberate action to minimize misregistration problems and ensure a high degree of confidence in the ground data labels. However, as a result of this type of strategy, the accuracy statement derived may be optimistically biased (Congalton & Plourde, 2000; Hammond & Verbyla, 1996; Muller et al., 1998; Yang et al., 2000; Zhu et al., 2000) and only relevant to a small part of the image. For example, Muller et al. (1998) estimated that the accuracy statement to accompany one classification derived was applicable to only 22% of the region mapped. To be applicable to the entire map, the sample used in forming the confusion matrix would have to be fully representative of the conditions found in the region (Congalton et al., 1998).

5.5. Spatial distribution of error

The erroneous allocations made by a classification are typically not randomly distributed over the region (Congalton, 1988; Steele, Winne, & Redmond, 1998). In the IGBP global land cover map, for example, the accuracy with which the individual continents are classified differs by ~20% (Loveland et al., 1999). Often there is a distinct pattern to the spatial distribution of thematic errors arising from the sensor's properties (Foody, 1988) and/or the ground conditions, with, for example, errors spatially correlated at the boundaries of classes (Congalton, 1988; Edwards & Lowell, 1996; Steele et al., 1998; Vieira &

Mather, 2000). Much of the error occurring at boundaries is associated with misregistration of the data sets and mixed pixels. Irrespective of their origin, the spatial variability of error can be a major concern, particularly in terms of error propagation. Unfortunately, however, the confusion matrix and the accuracy metrics derived from it provide no information on the spatial distribution of error (Canters, 1997; Morisette, Khorram, & Mace, 1999; Steele et al., 1998; Vaesen, Lizarraga, Nackaerts, & Coppin, 2000).

Many users of thematic maps derived from remotely sensed data may benefit from a spatial representation of classification accuracy. Various approaches have been investigated to provide this information. These include extrapolations from the training set (Steele et al., 1998), descriptors of the magnitude and partitioning of class membership among the classes (Foody, 2000a), and the use of a geostatistical approach to model variation in accuracy over the mapped region (Kyriakidis & Dungan, in press). Much of the recent effort on representing the spatial distribution of classification quality has been directed at the visualization of classification uncertainty (Fisher, 1994; Maselli, Conese, & Petkov, 1994; Thierry & Lowell, 2001). This has typically been derived as a by-product of the classification analysis. For example, the uncertainty in a conventional (hard) class allocation may be derived through examination of the posterior probabilities of class membership calculated in the course of a maximum likelihood classification (Canters, 1997; Foody, Campbell, Trodd, & Wood, 1992). This has often been used to provide an indication of the quality of the classification on a per-case basis to supplement the global summary provided by standard accuracy statements. In addition, this information may be used to indicate the spatial distribution of accuracy and so the location of problematic areas (Corves & Place, 1994; Foody, 2000a; Foody et al., 1992; Maselli et al., 1994; Van Deusen, 1995). Thus, the provision of an image depicting the spatial variation in classification uncertainty may be a useful accompaniment to the classification itself. However, it is important to recognize that the image pixels are not independent samples and that geostatistical techniques may be usefully employed in attempting to represent the spatial variation in classification uncertainty (de Bruin, 2000a).

5.6. Error magnitude

In classical accuracy assessments all misallocations are equally weighted. Often, however, some errors are more important or damaging than others (Forbes, 1995; Naesset, 1996; Stehman, 1999a). In many instances, the errors observed in a classification are between relatively similar classes (Felix & Binney, 1989; Loveland et al., 1999; Mickelson et al., 1998; Zhu et al., 2000) and sometimes these may be unimportant while other errors may be highly significant (DeFries & Los, 1999). Often, for example, error arises as a function of class definition, particularly through attempts to represent continua by a set of discrete classes (Felix & Binney, 1989; Foody, 2000a; Steele et al., 1998; Townsend, 2000). This can be a major source of classification error (Todd et al., 1980), much of which is associated with the allocation of relatively similar cases to different classes either side of (typically arbitrary) class boundaries (Campbell & Mortenson, 1989; Foody, 2000a; Sheppard et al., 1995). The conventional (hard) allocation of sites to discrete classes is, therefore, an issue of concern in thematic mapping, particularly as it may sometimes be more appropriate to model continuous variations in land surface properties and allow for indeterminate boundaries (Andrefouet & Roux, 1998; DeFries, Hansen, & Townshend, 2000; Foody, 1996). The assessment of the accuracy of these representations is, however, difficult (Foody, 1996; Townsend, 2000) and is discussed below. However, if a standard hard classification is used to represent classes that lie along a continuum, error may range in magnitude from relatively minor confusion involving similar classes either side of an arbitrarily defined class boundary to the confusion of the very dissimilar classes located at the end points of the continuum. The importance of these different degrees of error may vary depending on the user's requirements but, in many instances, the assumption that errors are of equal severity is untenable. This is unfortunate as basic accuracy metrics such as the percentage of cases correctly allocated or the kappa coefficient effectively weighed errors equally.

The utility of some measures of accuracy is limited by variations in the severity of error magnitude (Adams & Hand, 1999). Some accuracy assessment procedures may, however, be adapted to accommodate known differences in error severity (Stehman, 1999a). For example, the various possible thematic errors that can be encountered in a study may be assigned differing scores of severity and a weighted kappa coefficient derived (Foody, Palubinskas, Lucas, Curran, & Honzak, 1996; Naesset, 1996), although the selection of the weights is subjective and the approach problematic (Stehman, 1997a). An attractive alternative is to use accuracy measures based on information theory (Kew, 1996). Not only may these allow a broader range of problems to be tackled (Kew, 1996), including the handling of nonsquare confusion matrices in which the classification schemes used in the thematic map and ground data differ (Finn, 1993), but they can also weight error magnitudes and may generally be more acceptable than measures such as the kappa coefficient of agreement (Forbes, 1995).

A further source of error associated with the use of a standard (hard) classifier that allocates each pixel to a single class is the implicit assumption that the image is composed of pure pixels (Foody, 1996; Gong & Howarth, 1990). Unfortunately, remotely sensed data are often dominated by pixels that represent areas containing more than one class and these are a major problem in accuracy assessment (Foody, 1996, 1999; Karaska, Huguenin, Van Blaricom, & Savitsky, 1995). Indeed, in some studies, mixed pixels have been identified as the most important cause of misclassifi-

cation (e.g., CCRS, 1999) and a major contributor to the misestimation of land cover change (Skole & Tucker, 1993). For example, mixing at class boundaries was a major problem in the land cover map of Great Britain with error concentrated around boundaries. Removal of these boundary regions enabled the estimated accuracy of the map to rise from 46% to 71% (Fuller et al., 1994). Unfortunately, mixed pixels are common, especially in coarse spatial resolution data sets and/or where the land cover mosaic is complex (Campbell, 1996; Crapper, 1984). In a standard (hard) classification of data containing mixed pixels, the interpretation of the class allocations made is difficult as many of the errors observed may be only partial errors, as the pixel may represent an area that is partly comprised of the allocated class. Similarly, however, some of the apparently correct class allocations may be partly erroneous. As any hard allocation of a mixed pixel must to some extent be erroneous, the presence of mixed pixels is a major problem in the use of classification techniques for thematic mapping. Indeed, the mixed pixel problem may be one of the main reasons why some map producers do not adopt the commonly recommended methods of accuracy assessment. As many remotely sensed data sets are dominated by mixed pixels, the standard accuracy assessment measures such as the kappa coefficient, which assume implicitly that each of the testing samples is pure, are, therefore, often inappropriate for accuracy assessment in remote sensing (Foody, 1996; Karaska et al., 1995). The strict insistence on perfect agreement between the image classification and the ground data may, therefore, be inappropriate. Again, the producers of other maps, such as those depicting soils, recognize the complexity of the feature being mapped, the subjectivity of the mapping process, and the impracticality of being specific and so typically opt to represent some feature such as the dominant class rather than attempt anything more specific (e.g., Curtis, Courtney, & Trudgill, 1976).

Alternative approaches to the standard crisp class memberships commonly assumed in remote sensing may sometimes be desired. Thus, rather than make simple, and perhaps overly severe, correct/incorrect evaluations, the analyst may instead adopt a scale of error severity that may be used in the accuracy assessment (DeGloria et al., 2000; Gopal & Woodcock, 1994; Mickelson et al., 1998; Woodcock & Gopal, 2000). Methods based on the linguistic scale of Gopal and Woodcock (1994) seem to offer considerable potential here, particularly as a means of tolerating some degree of disagreement between the classification and ground data. If it is accepted that no one accuracy measure satisfies all users and that more than one measure should be provided along with a confusion matrix, it may be appropriate for one of these measures to be a fuzzy accuracy assessment measure and the other a standard accuracy metric as, depending on the metrics used, these can be very different but complementary measures of accuracy (Muller et al., 1998).

In attempting to solve the mixed pixel problem, fuzzy or soft classifications have been used increasingly. These typically are fuzzy in the sense that they allow each pixel to have multiple and partial class membership. These approaches have also proved popular in the representation of continua where the assumed conditions of discrete and mutually exclusive classes are unsatisfied (Foody, 1996; Townsend, 2000). Sometimes the uncertainty in making a conventional hard class allocation may be used to derive a fuzzy classification. For example, relative measures of class membership such as the posterior probabilities of class membership derived during a maximum likelihood classification may be used to indicate the composition of image pixels. Although the theoretical link between such measures of the uncertainty in making a particular allocation and subpixel composition is debatable (Canters, 1997; de Bruin, 2000b; Lewis, Nixon, Tatnall, & Brown, 1999), the uncertainty of the allocation is clearly a function of the pixel's composition and there are numerous examples that demonstrate the practical success of the method (e.g., Foody, 1996; Lewis et al., 1999). Alternatively, absolute measures of the strength of class membership may be used to represent continuous classes (Foody, 2000b). Irrespective of how derived, soft or fuzzy classifications cannot be sensibly evaluated in the normal way with a conventional confusion matrix-based analysis, even if the pixel's membership is partitioned among the classes in the matrix (as it is no longer really site-specific), and alternative measures have been sought. Frequently, such classifications have been evaluated using information theory-based metrics employing measures of entropy (e.g., Foody, 1996) or have been based on a linguistic scale of accuracy (Gopal & Woodcock, 1994). However, some recent work has sought to develop approaches founded on fuzzy and rough sets that are based on the confusion matrix (Ahlqvist, Keukelaar, & Oukbir, 2000; Binaghi, Brivio, Ghezzi, & Rampini, 1999; Jager & Benz, 2000; Lewis & Brown, in press; Matsakis, Andrefouet, & Calpolsini, 2000). Such approaches exploit the attractive features and popularity of the confusion matrix while usefully extending its applicability to fuzzy classifications. These and other approaches are important as the ability to evaluate the accuracy of fuzzy classifications will help to substantially reduce the effect of the mixed pixel problem. Since mixed pixels often dominate remotely sensed imagery and will not disappear with the use of fine spatial resolution data, techniques that allow their inclusion into the assessment of classification accuracy are required and there is scope for considerable research on this topic.

5.7. Use of the confusion matrix

The confusion matrix has been used mainly to provide a basic description of thematic map accuracy and for the comparison of accuracies. However, it may be possible to use the information contained in the matrix to derive considerably more useful information. As noted above, the confusion matrix may be useful in refining estimates of the areal extent of classes in the region. The confusion

matrix may, however, be used to further enhance the value of the classification for the user. In particular, it may be possible to use the matrix to help optimize the thematic map for a particular user (Lark, 1995; Morisette & Khorram, 2000). Thus, the matrix may be usefully employed with information on the actual costs of errors or the value of the map to optimize a classification for a particular application (Smits et al., 1999; Stehman, 1999a). Smits et al. (1999), for example, illustrate how a confusion matrix may be used together with information on the economic cost of misclassification to refine a thematic mapping investigation. In particular, it is shown that the results of such an analysis may be used to refocus the investigation or question the appropriateness of the data sets or methods used in deriving the classification. The utility of such methods, however, clearly depends on the reliability of the confusion matrix. Forming a reliable confusion matrix, in which one can be confident that issues discussed above (e.g., sample design, ground data accuracy, registration of the data sets etc.) have not had a detrimental effect is, however, difficult (Smits et al., 1999).

5.8. Accuracy of land cover change products

There is considerable interest in the use of remote sensing to study thematic change, such as land cover dynamics. This arises particularly through the importance of land cover change within the broader arena of environmental change (Skole, 1994) as well as the need to inform environmental policy and management decisions (Biging et al., 1999). Many methods of change detection have been used to study land cover change (Lambin & Ehrlich, 1997; Mas, 1999; Singh, 1989), but by far, the most popular has been the use of postclassification comparison methods.

A variety of factors influence the accuracy of land cover change products. With the popular postclassification comparison methods basic issues are the accuracies of the component classifications as well as more subtle issues associated with the sensors and data preprocessing methods used together with the prevailing conditions at the times of image acquisition (e.g., atmospheric properties, viewing geometry, etc.) (Khorram, 1999). In mapping land cover change, the problems noted above in relation to the registration of data sets and boundaries are generally magnified (Khorram, 1999; Roy, 2000). Error in the individual classifications may also be confused with change (Khorram, 1999). This can be difficult to allow for or study particularly as the location of boundaries between classes at each individual time period may be uncertain (De Groeve & Lowell, 2001) and there may be no information on the spatial distribution of accuracy for the classifications used. Consequently, any differences observed over time may not be attributable solely, if at all, to real change on the ground. As a consequence of these and other issues, the estimation of the accuracy of a change product is a substantially more difficult and challenging task than the assessment of the accuracy of a single image classification (Congalton & Green, 1999). With no standard approach to the assessment of the accuracy of a change product, it has been popular to adapt the standard confusion matrix to yield a change detection confusion matrix. The elements of this change detection confusion matrix represent individual from/to class change scenarios (Congalton & Green, 1999; Khorram, 1999). As a result, the dimensions of the matrix are much larger than the basic confusion matrix used to assess the accuracy of the single date classifications depicting the land cover classes of interest; each dimension of the change detection confusion matrix is the square of the number of classes involved. If desired, however, the matrix can be compressed into a 2×2 matrix illustrating simple change or no-change situations (Congalton & Green, 1999; Morisette & Khorram, 2000). From each type of matrix, some of the basic measures of accuracy discussed above can be derived to express the accuracy of the change detection (Biging et al., 1999; Congalton & Green, 1999). Obtaining the sample of data to use in the construction of the change detection confusion matrix can, however, be difficult. Often, for example, some of the change scenarios are rare, complicating the sampling process (Biging et al., 1999; Khorram, 1999). Perhaps a more significant problem, however, is that these approaches are appropriate only for use with conventional hard classifications. This, however, limits the change detection to indicating where a conversion of land cover appears to have occurred. Although land cover conversions are important, they are only one component of land cover change. Subtle transformations, land cover modifications, in which the land cover type may have been altered but not changed (e.g., a grassland degraded, a forest thinned), will be inappropriately represented by conventional postclassification comparison methods of change detection. This is unfortunate as land cover modifications may be as significant environmentally as land cover conversions (Foody, 2001a; Lambin, 1997). In general, the use of hard classifications within a postclassification comparison-based approach would be expected to underestimate the area of land undergoing a change and, where a change is detected, overestimate the magnitude of change as it is a simple binary technique (e.g., Foody, 2001a). This is a major limitation in environmental studies where the magnitude of change is often important. The ability to monitor land cover modifications associated with land degradation or rehabilitation would, for example, help inform environmental policy and decision making that underpin sustainable resource use (Foody, 2001b).

6. Summary and conclusions

Although thematic maps are an imperfect model of the environment, they are widely used and often derived from remotely sensed data through some form of classification analysis. The value of the map is clearly a function of the

accuracy of the classification. Unfortunately, the assessment of classification accuracy is not a simple task. Accuracy assessment in remote sensing has a long and, at times, contentious history. Accuracy assessment has, however, matured considerably and is now generally accepted to be a fundamental part of any thematic mapping exercise. Although there is no accepted standard method of accuracy assessment or reporting, the topic has matured to the extent that the general format of these important components of the mapping exercise can be identified. Typically, for example, the community is urged to base accuracy assessment on the confusion matrix and provide at least one quantitative metric of classification accuracy together with appropriate confidence limits. Additionally, some general level of accuracy is typically specified as a target against which the classification may be evaluated.

The confusion matrix lies at the core of much work on accuracy assessment and is frequently used without question to its suitability. The confusion matrix is used to provide a site-specific assessment of the correspondence between the image classification and ground conditions. The confusion matrix may, for example, be used to summarize the nature of the class allocations made by a classification and is the basis of many quantitative metrics of classification accuracy. However, there are many problems with accuracy assessment. A key concern is that the basic assumptions underlying the assessment of classification accuracy may not be satisfied. Rarely, for instance, will the data used be truly site-specific due to problems of mixed pixels and misregistration of the ground and remotely sensed data sets. The classes defined are also typically a generalization that may often be problematic. Moreover, rarely are the ground data an accurate representation of the ground conditions or the necessary information on the sampling design used in their acquisition provided. Obtaining a reliable confusion matrix is, therefore, a weak link in the accuracy assessment chain (Smits et al., 1999), yet it remains central to most accuracy assessment and reporting. Until basic problems such as those associated with mixed pixels and data set registration are solved, the interpretation of the confusion matrix and accuracy metrics derived from it will remain problematic.

Although there have been many recent advances, the current status of accuracy assessment indicates that numerous problems remain to be solved. Thus, although the subject has matured considerably, there is scope for significant further development. A key concern is that the widely used approaches for accuracy assessment and reporting are often flawed. Despite the apparent objectivity of quantitative metrics of accuracy, it is important that accuracy statements be interpreted with care. Many factors may result in a misleading interpretation being derived from an apparently objective accuracy statement. This situation could have serious implications for some users and may lessen their confidence in remote sensing as a source of land cover data. Considerable further research is therefore needed. This research needs to address basic image processing issues such as the registration of data sets as well as issues associated more closely with accuracy assessment itself. The latter may usefully benefit from investigations into defining tolerable levels of error and error reporting. In the meantime, accuracy assessment will remain a difficult issue. Given the range of viewpoints and problems, it seems unlikely that a single universally acceptable standard for accuracy assessment and reporting can be specified. What the remote sensing community can do, however, is provide more information than it currently seems to do. While the information required may vary depending on the aims of the mapping study and availability of the derived map to other users, a detailed account of the approach to accuracy assessment adopted may facilitate more informed use of the map than is otherwise possible. Thus, in addition to providing basic accuracy assessment components such as a quantitative metric of accuracy and a confusion matrix, if appropriate, it may be helpful if the information on issues such as the sampling design used to acquire the testing set, the confidence in the ground data labels, the classification protocols and lineage of the data sets used were provided. Many of these issues have been raised in the literature before and have been simply revisited here with reference to recent examples, which act to emphasize the need for the community to critically evaluate its procedures and change as appropriate in order to progress. This is particularly important with regard to the mapping of very large areas and monitoring of change, where accuracy assessment remains a challenging task with considerable scope for further development. It is, however, more difficult to gain support for research on classification accuracy assessment than classification projects (Scepan, 1999), yet the value of such projects will be limited ultimately by the poor quality of the accuracy assessment and reporting.

Acknowledgments

This article is based on a keynote address made at the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences (Accuracy 2000) symposium in Amsterdam in July 2000. I am grateful to the three referees for their highly constructive comments on the original manuscript.

References

Abrams, M., Bianchi, R., & Pieri, D. (1996). Revised mapping of lava flows on Mount Etna, Sicily. *Photogrammetric Engineering and Remote Sensing*, *62*, 1353–1359.

Adams, N. M., & Hand, D. J. (1999). Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition*, *32*, 1139–1147.

Ahlqvist, O., Keukelaar, J., & Oukbir, K. (2000). Rough classification and accuracy assessment. *International Journal of Geographical Information Science*, *14*, 475–496.

Anderson, J. R., Hardy, E. E., Roach, J. T., & Witmer, R. E. (1976). *A land*

use and land cover classification system for use with remote sensor data. Washington, DC: Government Printing Office (US Geological Survey, Professional Paper 964).

Andrefouet, S., & Roux, L. (1998). Characterisation of ecotones using membership degrees computed with a fuzzy classifier. *International Journal of Remote Sensing*, *19*, 3205–3211.

Arbia, G., Griffith, D., & Haining, R. (1998). Error propagation modelling in raster GIS: overlay operations. *International Journal of Geographical Information Science*, *12*, 145–167.

Aronoff, S. (1982). Classification accuracy: a user approach. *Photogrammetric Engineering and Remote Sensing*, *48*, 1299–1307.

Aronoff, S. (1985). The minimum accuracy value as an index of classification accuracy. *Photogrammetric Engineering and Remote Sensing*, *51*, 99–111.

Atkinson, P. M. (1991). Optimal ground-based sampling for remote sensing investigations: estimating the regional mean. *International Journal of Remote Sensing*, *12*, 559–567.

Atkinson, P. M., Foody, G. M., Curran, P. J., & Boyd, D. S. (2000). Assessing the ground data requirements for regional-scale remote sensing of tropical forest biophysical properties. *International Journal of Remote Sensing*, *21*, 2571–2587.

Bastin, L., Edwards, M., & Fisher, P. (2000). Tracking the positional uncertainty in 'ground truth'. In: G. B. M. Heuvelink, M. J. P. M. Lemmens (Eds.), *Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences* (pp. 39–42). Delft: Delft University Press.

Bauer, M. E., Burk, T. E., Ek, A. R., Coppin, P. R., Lime, S. D., Walsh, T. A., & Walters, D. K. (1994). Satellite inventory of Minnesota forest resources. *Photogrammetric Engineering and Remote Sensing*, *60*, 287–298.

Belward, A. S., Estes, J. E., & Kilne, K. D. (1999). The IGBP-DIS global 1-km land-cover data set DISCover: a project overview. *Photogrammetric Engineering and Remote Sensing*, *65*, 1013–1020.

Biging, G. S., Colby, D. R., & Congalton, R. G. (1999). Sampling systems for change detection accuracy assessment. In: R. S. Lunetta, & C. D. Elvidge (Eds.), *Remote sensing change detection: environmental monitoring methods and applications* (pp. 281–308). London: Taylor and Francis.

Binaghi, E., Brivio, P. A., Ghezzi, P., & Rampini, A. (1999). A fuzzy set-based accuracy assessment of soft classification. *Pattern Recognition Letters*, *20*, 935–948.

Bird, A. C., Taylor, J. C., & Brewer, T. R. (2000). Mapping National Park landscape from ground, air and space. *International Journal of Remote Sensing*, *21*, 2719–2736.

Bowers, T. L., & Rowan, L. C. (1996). Remote mineralogic and lithologic mapping of the Ice River Alkaline Complex, British Columbia, Canada using AVIRIS data. *Photogrammetric Engineering and Remote Sensing*, *62*, 1379–1385.

Brown, J. F., Loveland, T. R., Ohlen, D. O., & Zhu, Z. (1999). The global land-cover characteristics database: the user's perspective. *Photogrammetric Engineering and Remote Sensing*, *65*, 1069–1074.

Buckton, D., O'Mongain, E., & Danaher, S. (1999). The use of neural networks for the estimation of oceanic constituents based on the MERIS instrument. *International Journal of Remote Sensing*, *20*, 1841–1851.

Campbell, J. B. (1996). *Introduction to remote sensing* (2nd ed.). London: Taylor and Francis.

Campbell, W. G., & Mortenson, D. C. (1989). Ensuring the quality of geographic information system data: a practical application of quality control. *Photogrammetric Engineering and Remote Sensing*, *55*, 1613–1618.

Canters, F. (1997). Evaluating the uncertainty of area estimates derived from fuzzy land-cover classification. *Photogrammetric Engineering and Remote Sensing*, *63*, 403–414.

CCRS (1999). New land cover map of Canada. *Remote Sensing in Canada, Canada Centre for Remote Sensing Newsletter*, *27*, 7–9.

Chapin, F. S. III, Zavaleta, E. S., Eviner, V. T., Naylor, R. L., Vitousek, P. M., Reynolds, H. L., Hooper, D. U., Lavorel, S., Sala, O. E., Hobbie, S. E., Mack, M. C., & Diaz, S. (2000). Consequences of changing biodiversity. *Nature*, *405*, 234–242.

Cihlar, J. (2000). Land cover mapping of large areas from satellites: status and research priorities. *International Journal of Remote Sensing*, *21*, 1093–1114.

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

Cohen, W. B., & Justice, C. O. (1999). Validating MODIS terrestrial ecology products: linking in situ and satellite measurements. *Remote Sensing of Environment*, *70*, 1–3.

Congalton, R. G. (1988). Using spatial autocorrelation analysis to explore the errors in maps generated from remotely sensed data. *Photogrammetric Engineering and Remote Sensing*, *54*, 587–592.

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

Congalton, R. G. (1994). Accuracy assessment of remotely sensed data: future needs and directions. In: *Proceedings of Pecora 12 land information from space-based systems* (pp. 383–388). Bethesda: ASPRS.

Congalton, R. G., Balogh, M., Bell, C., Green, K., Milliken, J. A., & Ottman, R. (1998). Mapping and monitoring agricultural crops and other land cover in the Lower Colorado river basin. *Photogrammetric Engineering and Remote Sensing*, *64*, 1107–1113.

Congalton, R. G., & Green, K. (1993). A practical look at the sources of confusion in error matrix generation. *Photogrammetric Engineering and Remote Sensing*, *59*, 641–644.

Congalton, R. G., & Green, K. (1999). *Assessing the accuracy of remotely sensed data: principles and practices*. Boca Raton: Lewis Publishers.

Congalton, R. G., Green, K., & Teply, J. (1993). Mapping old growth forests on National Forest and Park Lands in the Pacific northwest from remotely sensed data. *Photogrammetric Engineering and Remote Sensing*, *59*, 529–535.

Congalton, R. G., & Plourde, L. C. (2000). Sampling methodology, sample placement, and other important factors in assessing the accuracy of remotely sensed forest maps. In: G. B. M. Heuvelink, M. J. P. M. Lemmens (Eds.), *Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences* (pp. 117–124). Delft: Delft University Press.

Corves, C., & Place, C. J. (1994). Mapping the reliability of satellite-derived landcover maps—an example from central Brazilian Amazon Basin. *International Journal of Remote Sensing*, *15*, 1283–1294.

Crapper, P. F. (1984). An estimate of the number of boundary cells in a mapped landscape coded to grid cells. *Photogrammetric Engineering and Remote Sensing*, *50*, 1497–1503.

Curtis, L. F., Courtney, F. M., & Trudgill, S. (1976). *Soils in the British Isles*. London: Longman.

Czaplewski, R. L. (1992). Misclassification bias in areal estimates. *Photogrammetric Engineering and Remote Sensing*, *58*, 189–192.

de Bruin, S. (2000a). Spatial uncertainty in estimates of the areal extent of land cover types. In: G. B. M. Heuvelink, M. J. P. M. Lemmens (Eds.), *Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences* (pp. 137–144). Delft: Delft University Press.

de Bruin, S. (2000b). Querying probabilistic land cover data using fuzzy set theory. *International Journal of Geographical Information Science*, *14*, 359–372.

DeFries, R. S., Hansen, M. C., & Townshend, J. R. G. (2000). Global continuous fields of vegetation characteristics: a linear mixture model applied to multi-year 8km AVHRR data. *International Journal of Remote Sensing*, *21*, 1389–1414.

DeFries, R. S., Hansen, M., Townshend, J. R. G., & Sohlberg, R. (1998). Global land cover classification at 8km spatial resolution: use of training data derived from Landsat imagery in decision tree classifiers. *International Journal of Remote Sensing*, *19*, 3141–3168.

DeFries, R. S., & Los, S. O. (1999). Implications of land-cover misclassification for parameter estimates in global land-surface models: an example from the simple biosphere model (SiB2). *Photogrammetric Engineering and Remote Sensing*, *65*, 1083–1088.

DeFries, R. S., & Townshend, J. R. G. (1994). Global land cover: comparison of ground-based data sets to classifications with AVHRR data. In: G. M. Foody, & P. J. Curran (Eds.), *Environmental remote sensing from regional to global scales* (pp. 84–110). Chichester: Wiley.

DeGloria, S. D., Laba, M., Gregory, S. K., Braden, J., Ogurcak, D., Hill, E., Fegraus, E., Fiore, J., Stalter, A., Beecher, J., Elliot, R., & Weber, J. (2000). Conventional and fuzzy accuracy assessment of land cover maps at regional scale. In: G. B. M. Heuvelink, M. J. P. M. Lemmens (Eds.), *Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences* (pp. 153–160). Delft: Delft University Press.

De Groeve, T., & Lowell, K. (2001). Boundary uncertainty assessment from a single forest-type map. *Photogrammetric Engineering and Remote Sensing*, *67*, 717–726.

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

Douglas, I. (1999). Hydrological investigations of forest disturbance and land cover impacts in South-East Asia: a review. *Philosophical Transactions of the Royal Society of London, Series B*, *354*, 1725–1738.

Edwards, G., & Lowell, K. E. (1996). Modeling uncertainty in photointerpreted boundaries. *Photogrammetric Engineering and Remote Sensing*, *62*, 377–391.

Edwards, T. C., Moisen, G. G., & Cutler, D. R. (1998). Assessing map accuracy in a remotely sensed, ecoregion-scale cover map. *Remote Sensing of Environment*, *63*, 73–83.

Estes, J., Belward, A., Loveland, T., Scepan, J., Strahler, A., Townshend, J., & Justice, C. (1999). The way forward. *Photogrammetric Engineering and Remote Sensing*, *65*, 1089–1093.

Estes, J. E., & Mooneyhan, D. W. (1994). Of maps and myths. *Photogrammetric Engineering and Remote Sensing*, *60*, 517–524.

Felix, N. A., & Binney, D. L. (1989). Accuracy assessment of a Landsat-assisted vegetation map of the coastal plain of the Arctic National Wildlife Refuge. *Photogrammetric Engineering and Remote Sensing*, *55*, 475–478.

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

Fisher, P. F. (1994). Visualization of the reliability in classified remotely sensed images. *Photogrammetric Engineering and Remote Sensing*, *60*, 905–910.

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

Foody, G. M. (1988). The effects of viewing geometry on image classification. *International Journal of Remote Sensing*, *9*, 1909–1915.

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. (1996). Approaches for the production and evaluation of fuzzy land cover classification from remotely-sensed data. *International Journal of Remote Sensing*, *17*, 1317–1340.

Foody, G. M. (1999). The continuum of classification fuzziness in thematic mapping. *Photogrammetric Engineering and Remote Sensing*, *65*, 443–451.

Foody, G. M. (2000a). Mapping land cover from remotely sensed data with a softened feedforward neural network classification. *Journal of Intelligent and Robotic Systems*, *29*, 433–449.

Foody, G. M. (2000b). Estimation of sub-pixel land cover composition in the presence of untrained classes. *Computers and Geosciences*, *26*, 469–478.

Foody, G. M. (2001a). Monitoring the magnitude of land-cover change around the southern limits of the Sahara. *Photogrammetric Engineering and Remote Sensing*, *67*, 841–847.

Foody, G. M. (2001b). Remote sensing of tropical forest environments: towards the monitoring of sustainable resource use. In: A. Belward, E. Binaghi, P. A. Brivio, G. A. Lanzarone, G. Tosi (Eds.), *Proceedings geo-spatial knowledge processing for natural resource management* (pp. 97–101). Varese, Italy: University of Insubria (28–29 June 2001).

Foody, G. M., Campbell, N. A., Trodd, N. M., & Wood, T. F. (1992). Derivation and applications of probabilistic measures of class membership from the maximum likelihood classification. *Photogrammetric Engineering and Remote Sensing*, *58*, 1335–1341.

Foody, G. M., Palubinskas, G., Lucas, R. M., Curran, P. J., & Honzak, M. (1996). Identifying terrestrial carbon sinks: classification of successional stages in regenerating tropical forest from Landsat TM data. *Remote Sensing of Environment*, *55*, 205–216.

Forbes, A. D. (1995). Classification-algorithm evaluation: five performance measures based on confusion matrices. *Journal of Clinical Monitoring*, *11*, 189–206.

Friedl, M. A., Woodcock, C., Gopal, S., Muchoney, D., Strahler, A. H., & Barker-Schaaf, C. (2000). A note on procedures used for accuracy assessment in land cover maps derived from AVHRR data. *International Journal of Remote Sensing*, *21*, 1073–1077.

Fuller, R. M., Groom, G. B., & Jones, A. R. (1994). The land cover map of Great Britain: an automated classification of Landsat Thematic Mapper data. *Photogrammetric Engineering and Remote Sensing*, *60*, 553–562.

Gong, P., & Howarth, P. J. (1990). The use of structural information for improving land-cover classification accuracies at the rural–urban fringe. *Photogrammetric Engineering and Remote Sensing*, *56*, 67–73.

Gopal, S., & Woodcock, C. (1994). Theory and methods for accuracy assessment of thematic maps using fuzzy sets. *Photogrammetric Engineering and Remote Sensing*, *60*, 81–188.

Green, E. J., Strawderman, W. E., & Airola, T. M. (1993). Assessing classification probabilities for thematic maps. *Photogrammetric Engineering and Remote Sensing*, *59*, 635–639.

Hammond, T. O., & Verbyla, D. L. (1996). Optimistic bias in classification accuracy assessment. *International Journal of Remote Sensing*, *17*, 1261–1266.

Hansen, M. C., DeFries, R. S., Townshend, J. R. G., & Sohlberg, R. (2000). Global land cover classification at 1km spatial resolution using a classification tree approach. *International Journal of Remote Sensing*, *21*, 1331–1364.

Hay, A. M. (1979). Sampling designs to test land-use map accuracy. *Photogrammetric Engineering and Remote Sensing*, *45*, 529–533.

Hay, A. M. (1988). The derivation of global estimates from a confusion matrix. *International Journal of Remote Sensing*, *9*, 1395–1398.

Husak, G. J., Hadley, B. C., & McGwire, K. C. (1999). Landsat Thematic Mapper registration accuracy and its effects on the IGBP validation. *Photogrammetric Engineering and Remote Sensing*, *65*, 1033–1039.

Jager, G., & Benz, U. (2000). Measures of classification accuracy based on fuzzy similarity. *IEEE Transactions on Geoscience and Remote Sensing*, *38*, 1462–1467.

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

Jensen, J. R. (1996). Introductory digital image processing. *A remote sensing perspective* (2nd ed.). New Jersey: Prentice-Hall.

Johnston, D. M., & Timlin, D. (2000). Spatial data accuracy and quality assessment for environmental management. In: G. B. M. Heuvelink, M. J. P. M. Lemmens (Eds.), *Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences* (pp. 325–328). Delft: Delft University Press.

Jupp, D. L. B. (1989). The stability of global estimates from confusion matrices. *International Journal of Remote Sensing*, *10*, 1563–1569.

Justice, C., Belward, A., Morisette, J., Lewis, P., Privette, J., & Baret, F. (2000). Developments in the 'validation' of satellite sensor products for the study of the land surface. *International Journal of Remote Sensing*, *21*, 3383–3390.

Kalkhan, M. A., Reich, R. M., & Czaplewski, R. L. (1995). Statistical properties of five indices in assessing the accuracy of remotely sensed data using simple random sampling. *Proceedings ACSM/ASPRS Annual Convention and Exposition*, *2*, 246–257.

Kalkhan, M. A., Reich, R. M., & Stohlgren, T. J. (1998). Assessing the accuracy of Landsat Thematic Mapper classification using double sampling. *International Journal of Remote Sensing*, *19*, 2049–2060.

Karaska, M. A., Huguenin, R. L., Van Blaricom, D., & Savitsky, B. (1995). Subpixel classification of cypress and tupelo trees in TM imagery. *Proceedings of the 1995 ACSM/ASPRS Annual Convention and Exposition*, *3*, 856–865.

Kew, N. R. (1996). Information-theoretic measures for assessment and analysis in image classification. In: E. Binaghi, P. A. Brivio, & A. Rampini (Eds.), *Soft computing in remote sensing data analysis* (pp. 173–180). Singapore: World Scientific.

Khorram, S. (Ed.) (1999). *Accuracy assessment of remote sensing-derived change detection*. Bethesda, MD: American Society for Photogrammetry and Remote Sensing.

Koukoulas, S., & Blackburn, G. A. (2001). Introducing new indices for accuracy evaluation of classified images represnting semi-natural woodland environments. *Photogrammetric Engineering and Remote Sensing*, *67*, 499–510.

Kyriakidis, P. C., & Dungan, J. L. (2001). A geostatistical approach for mapping thematic classification accuracy and evaluating the impact of inaccurate spatial data on ecological model predictions. *Environmental and Ecological Statistics*, (in press).

Lambin, E. F. (1997). Modelling and monitoring land-cover change processes in tropical regions. *Progress in Physical Geography*, *21*, 375–393.

Lambin, E. F., & Ehrlich, D. (1997). Land-cover changes in sub-Saharan Africa (1982–1991): application of a change index based on remotely sensed surface temperature and vegetation indices at a continental scale. *Remote Sensing of Environment*, *61*, 181–200.

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

Lewis, H. G., & Brown, M. (2001). A generalised confusion matrix for assessing area estimates from remotely sensed data. *International Journal of Remote Sensing*, (in press).

Lewis, H. G., Nixon, M. S., Tatnall, A. R. L., & Brown, M. (1999). Appropriate stategies for mapping land cover from satellite imagery. In: *Proceedings of RSS99 earth observation, from data to information* (pp. 717–724). Nottingham: Remote Sensing Society.

Loveland, T. R., Reed, B. C., Brown, J. F., Ohlen, D. O., Zhu, Z., Yang, L., & Merchant, J. W. (2000). Development of a global land cover characteristics database and IGBP DISCover from 1km AVHRR data. *International Journal of Remote Sensing*, *21*, 1303–1330.

Loveland, T. R., Zhu, Z., Ohlen, D. O., Brown, J. F., Reed, B. C., & Yang, L. (1999). An analysis of the IGBP global land-cover characterisation process. *Photogrammetric Engineering and Remote Sensing*, *65*, 1021–1032.

Lunetta, R. S., Iiames, J., Knight, J., Congalton, R. G., & Mace, T. H. (2001). An assessment of reference data variability using a ''virtual field reference database''. *Photogrammetric Engineering and Remote Sensing*, *63*, 707–715.

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

Maling, D. H. (1989). *Measurements from maps*. Oxford: Pergamon.

Mas, J.-F. (1999). Monitoring land-cover changes: a comparison of change detection techniques. *International Journal of Remote Sensing*, *20*, 139–152.

Maselli, F., Conese, C., & Petkov, L. (1994). Use of probability entropy for the estimation and graphical representation of the accuracy of maximum likelihood classifications. *ISPRS Journal of Photogrammetry and Remote Sensing*, *49*, 13–20.

Mather, P. M. (1999). Land cover classification revisited. In: P. M. Tate, & N. J. Tate (Eds.), *Advances in remote sensing and GIS analysis* (pp. 7–16). Chichester: Wiley.

Matsakis, P., Andrefouet, S., & Capolsini, P. (2000). Evaluation of fuzzy partitions. *Remote Sensing of Environment*, *74*, 516–533.

Merchant, J. W., Yang, L., & Yang, W. (1994). Validation of continental-scale land cover data bases developed from AVHRR data. In: *Proceedings of Pecora 12 land information from space-based systems* (pp. 63–72). Bethesda: ASPRS.

Mickelson, J. G., Civco, D. L., & Silander, J. A. (1998). Delineating forest canopy species in the Northeastern United States using multi-temporal TM imagery. *Photogrammetric Engineering and Remote Sensing*, *64*, 891–904.

Morisette, J. T., & Khorram, S. (2000). Accuracy-assessment curves for satellite-based change detection. *Photogrammetric Engineering and Remote Sensing*, *66*, 875–880.

Morisette, J. T., Khorram, S., & Mace, T. (1999). Land-cover change detection enhanced with generalized linear models. *International Journal of Remote Sensing*, *20*, 2703–2721.

Muchoney, D., Borak, J., Chi, H., Friedl, M., Gopal, S., Hodges, N., Morrow, N., & Strahler, A. (2000). Applications of the MODIS global supervised classification model to vegetation and land cover mapping of Central America. *International Journal of Remote Sensing*, *21*, 1115–1138.

Muchoney, D., Strahler, A., Hodges, J., & LoCastro, J. (1999). The IGBP DISCover confidence sites and the system for terrestrial ecosystem parameterization: tools for validating global land-cover data. *Photogrammetric Engineering and Remote Sensing*, *65*, 1061–1067.

Muller, S. V., Walker, D. A., Nelson, F. E., Auerach, N. A., Bockheim, J. G., Guyer, S., & Sherba, D. (1998). Accuracy assessment of a land-cover map of the Kuparuk river basin, Alaska: considerations for remote regions. *Photogrammetric Engineering and Remote Sensing*, *64*, 619–628.

Naesset, E. (1996). Use of weighted kappa coefficient in classification error assessment of thematic maps. *International Journal of Geographical Information Systems*, *10*, 591–604.

Penner, J. E. (1994). Atmospheric chemistry and air quality. In: W. B. Meyer, B. L. Turner II (Eds.), *Changes in land use and land cover: a global perspective* (pp. 175–209). Cambridge: Cambridge University Press.

Piper, S. E. (1983). The evaluation of the spatial accuracy of computer classification. In: *Proceedings of the 1983 Machine Processing of Remotely Sensed Data Symposium* (pp. 303–310). West Lafayette: Purdue University.

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

Prisley, S. P., & Smith, J. L. (1987). Using classification error matrices to improve the accuracy of weighted land-cover models. *Photogrammetric Engineering and Remote Sensing*, *53*, 1259–1263.

Purvis, A., & Hector, A. (2000). Getting the measure of biodiversity. *Nature*, *405*, 212–219.

Rhind, D., & Hudson, R. (1980). *Land use*. London: Methuen.

Richards, J. A. (1996). Classifier performance and map accuracy. *Remote Sensing of Environment*, *57*, 161–166.

Riemann, R., Hoppus, M., & Lister, A. (2000). Using arrays of small ground sample plots to assess the accuracy of Landsat TM-derived forest-cover maps. In: G. B. M. Heuvelink, M. J. P. M. Lemmens (Eds.), *Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences* (pp. 541–548). Delft: Delft University Press.

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

Roy, D. P. (2000). The impact of misregistration upon composited wide field of view satellite data and implications for change detection. *IEEE Transactions on Geoscience and Remote Sensing*, *38*, 2017–2032.

Scepan, J. (1999). Thematic validation of high-resolution global land-cover data sets. *Photogrammetric Engineering and Remote Sensing*, *65*, 1051–1060.

Scepan, J., Menz, G., & Hansen, M. C. (1999). The DISCover validation image interpretation process. *Photogrammetric Engineering and Remote Sensing*, *65*, 1075–1081.

Schlagel, J. D., & Newton, C. M. (1996). A GIS-based statistical method to

analyze spatial change. *Photogrammetric Engineering and Remote Sensing*, *62*, 839–844.

Sheppard, C. R. C., Matheson, K., Bythell, J. C., Murphy, P., Myers, C. B., & Blake, B. (1995). Habitat mapping in the Caribbean for management and conservation: use and assessment of aerial photography. *Aquatic Conservation: Marine and Freshwater Ecosystems*, *5*, 277–298.

Singh, A. (1989). Digital change detection techniques using remotely-sensed data. *International Journal of Remote Sensing*, *10*, 989–1003.

Skole, D., & Tucker, C. (1993). Tropical deforestation and habitat fragmentation in the Amazon: satellite data from 1978 to 1988. *Science*, *260*, 1905–1910.

Skole, D. L. (1994). Data on global land-cover change: acquisition, assessment and analysis. In: W. B. Meyer, B. L. Turner II (Eds.), *Changes in land use and land cover: a global perspective* (pp. 437–471). Cambridge: Cambridge University Press.

Smedes, H. W. (1975). The truth about ground truth. In: *Proceedings 10th International Symposium on Remote Sensing of Environment* (pp. 821–823). Ann Arbor, Michigan: Environmental Research Institute of Michigan.

Smits, P. C., Dellepiane, S. G., & 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*, 1461–1486.

Steele, B. M., Winne, J. C., & Redmond, R. L. (1998). Estimation and mapping of misclassification probabilities for thematic land cover maps. *Remote Sensing of Environment*, *66*, 192–202.

Stehman, S. V. (1995). Thematic map accuracy assessment from the perspective of finite population sampling. *International Journal of Remote Sensing*, *16*, 589–593.

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

Stehman, S. V. (1996b). Use of auxiliary data to improve the precision of estimators of thematic map accuracy. *Remote Sensing of Environment*, *58*, 169–176.

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

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

Stehman, S. V. (1999a). Comparing thematic maps based on map value. *International Journal of Remote Sensing*, *20*, 2347–2366.

Stehman, S. V. (1999b). Basic probability sampling designs for thematic map accuracy assessment. *International Journal of Remote Sensing*, *20*, 2423–2441.

Stehman, S. V. (2000). Practical implications of deign-based sampling for thematic map accuracy assessment. *Remote Sening of Environment*, *72*, 35–45.

Stehman, S. V. (2001). Statistical rigor and practical utility in thematic map accuracy assessment. *Photogrammetric Engineering and Remote Sensing*, *67*, 727–734.

Stehman, S. V., & Czaplewski, R. L. (1998). Design and analysis for thematic map accuracy assessment: fundamental principles. *Remote Sensing of Environment*, *64*, 331–344.

Stehman, S. V., Wickham, J. D., Yang, L., & Smith, J. H. (2000). Assessing the accuracy of large-area land cover maps: experiences from the multi-resolution land-cover charactersitics (MRLC) project. In: G. B. M. Heuvelink, M. J. P. M. Lemmens (Eds.), *Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences* (pp. 601–608). Delft: Delft University Press.

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

Thierry, B., & Lowell, K. (2001). An uncertainty-based method of photo-interpretation. *Photogrammetric Engineering and Remote Sensing*, *67*, 65–72.

Thomas, I. L., & Allcock, G. McK. (1984). Determining the confidence level for a classification. *Photogrammetric Engineering and Remote Sensing*, *50*, 1491–1496.

Thomlinson, J. R., Bolstad, P. V., & Cohen, W. B. (1999). Coordinating methodologies for scaling landcover classifications from site-specific to global: steps toward validating global map products. *Remote Sensing of Environment*, *70*, 16–28.

Todd, W. J., Gehring, D. G., & Haman, J. F. (1980). Landsat wildland mapping accuracy. *Photogrammetric Engineering and Remote Sensing*, *46, 509–520.*

Townsend, P. A. (2000). A quantitative fuzzy approach to assess mapped vegetation classification for ecological applications. *Remote Sensing of Environment*, *72*, 253–267.

Townshend, J. R. G. (1992). Land cover. *International Journal of Remote Sensing*, *13*, 1319–1328.

Trodd, N. M. (1995). Uncertainty in land cover mapping for modelling land cover change. In: *Proceedings of RSS95 remote sensing in action* (pp. 1138–1145). Nottingham: Remote Sensing Society.

Turk, G. (1979). GT index: a measure of the success of prediction. *Remote Sensing of Environment*, *8*, 75–86.

Ung, C.-H., Lambert, M.-C., Guidon, L., & Fournier, R. A. (2000). Integrating Landsat-TM data with environmental data for classifying forest cover types and estimating their biomass. In: G. B. M. Heuvelink, M. J. P. M. Lemmens (Eds.), *Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences* (pp. 659–662). Delft: Delft University Press.

Vaesen, K., Lizarraga, I., Nackaerts, K., & Coppin, P. (2000). Spatial characterisation of uncertainty in forest change detection. In: G. B. M. Heuvelink, M. J. P. M. Lemmens (Eds.), *Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences* (pp. 671–674). Delft: Delft University Press.

Van Deusen, P. C. (1995). Modified highest confidence first classification. *Photogrammetric Engineering and Remote Sensing*, *61*, 419–425.

Van Deusen, P. C. (1996). Unbiased estimates of class proportions from thematic maps. *Photogrammetric Engineering and Remote Sensing*, *62*, 409–412.

Veregin, H. (1994). Integration of simulation modeling and error propagation for the buffer operation in GIS. *Photogrammetric Engineering and Remote Sensing*, *60*, 427–435.

Vieira, C. A. O., & Mather, P. M. (2000). Visulisation of measures of classifier reliability and error in remote sensing. In: G. B. M. Heuvelink, M. J. P. M. Lemmens (Eds.), *Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences* (pp. 701–708). Delft: Delft University Press.

Vitousek, P. M. (1994). Beyond global warming: ecology and global change. *Ecology*, *75*, 1861–1876.

Wickham, J. D., O'Neill, R. V., Ritters, K. H., Wade, T. G., & Jones, K. B. (1997). Sensitivity of selected landscape pattern metrics to land-cover misclassification and differences inland-cover composition. *Photogrammetric Engineering and Remote Sensing*, *63*, 397–402.

Wilkinson, G. G. (1996). Classification algorithms—where next? In: E. Brivio, P. A. Brivio, & A. Rampini (Eds.), *Soft computing in remote sensing data analysis* (pp. 93–99). Singapore: World Scientific.

Woodcock, C. E., & Gopal, S. (2000). Fuzzy set theory and thematic maps: accuracy assessment and area estimation. *International Journal of Geographical Information Science*, *14*, 153–172.

Worboys, M. (1998). Imprecision in finite resolution spatial data. *GeoInformatica*, *2*, 257–279.

Yang, L., Stehman, S. V., Wickham, J., Jonathan, S., & VanDriel, N. J. (2000). Thematic validation of land cover data of the eastern United States using aerial photography: feasibility and challenges. In: G. B. M. Heuvelink, M. J. P. M. Lemmens (Eds.), *Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences* (pp. 747–754). Delft: Delft University Press.

Yuan, D. (1997). A simulation comparison of three marginal area estimators

for image classification. *Photogrammetric Engineering and Remote Sensing*, *63*, 385–392.

Zhang, J., & Foody, G. M. (1998). A fuzzy classification of sub-urban land cover from remotely sensed imagery. *International Journal of Remote Sensing*, *19*, 2721–2738.

Zhou, Q., Robson, M., & Pilesjo, P. (1998). On the ground estimation of vegetation cover in Australian rangelands. *International Journal of Remote Sensing*, *19*, 1815–1820.

Zhu, Z., Yang, L., Stehman, S. V., & Czaplewski, R. L. (2000). Accuracy assessment for the U.S. Geological Survey regional land-cover mapping programme: New York and New Jersey region. *Photogrammetric Engineering and Remote Sensing*, *66*, 1425–1435.