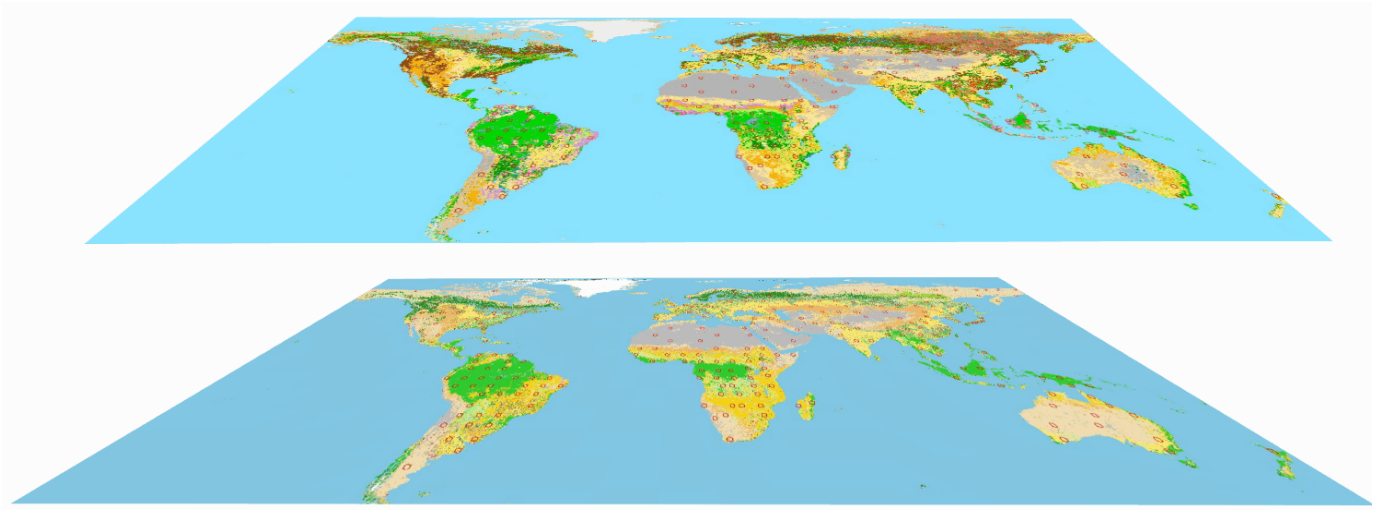
GLOBAL LAND COVER VALIDATION:

RECOMMENDATIONS FOR EVALUATION AND ACCURACY ASSESSMENT OF GLOBAL LAND COVER MAPS



GOFC-GOLD GLOBAL OBSERVATION OF FOREST AND LAND COVER DYNAMICS

EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

2006 EUR 22156 EN

GLOBAL LAND COVER VALIDATION:

RECOMMENDATIONS FOR EVALUATION AND ACCURACY ASSESSMENT OF GLOBAL LAND COVER MAPS

By

Alan H. Strahler[1], Luigi Boschetti[2,6], Giles M. Foody[3], Mark A. Friedl[1], Matthew C. Hansen[4], Martin Herold[5], Philippe Mayaux[6], Jeffrey T. Morisette[7], Stephen V. Stehman[8] and Curtis E. Woodcock[1]

[1] Boston University, Boston, USA

[2] University of Maryland, College Park, USA

[3] University of Southampton, United Kingdom

[4] South Dakota State University, Brookings, USA

[5] Friedrich Schmiller Universität, Jena, Germany

[6] Joint Research Centre of the European Commission, Ispra, Italy

[7] NASA Goddard Space Flight Center, Greenbelt, USA

[8] State University of New York, Syracuse, USA

LEGAL NOTICE

Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use which might be made of the following information.

A great deal of additional information on the European Union is available on the Internet. It can be accessed through the Europa server (http://europa.eu.int)

Luxembourg: Office for Official Publications of the European Communities, 2006

Printed in Italy

Global Land Cover Validation:
Recommendations for Evaluation and Accuracy Assessment
Of Global Land Cover Maps

Contents

Global Land Cover Validation:
Recommendations for Evaluation and Accuracy Assessment
Of Global Land Cover Maps

1. Introduction

This document presents the findings of the Working Group on Global Land Cover Validation, a topical group within the Land Product Validation (LPV) subgroup of the Working Group on Calibration and Validation (WGCV) of the Committee on Earth Observing Satellites (CEOS). The Land Product Validation subgroup (LPV) was established in 2000 with the mission to foster quantitative validation of higher-level global land products derived from remote sensing data and provide quantitative results that are relevant to users. It responds to a concerted international effort to assure the validation of global land science data products now being made available from a new generation of satellite sensors. To meet this mission, the LPV subgroup has established the following objectives:

- Work with users to define uncertainty objectives
- Identify and support global test sites for both systematic and episodic measurements
- Identify opportunities for coordination and collaboration
- Develop consensus "best practice" protocols for data collection and description
- Develop procedures for validation data exchange and management (see http://landval.gsfc.nasa.gov/LPVS)

Our report represents a major contribution to the fourth objective above for global land cover products.

This document arises from presentations, discussions, and group activities at two workshops devoted to validation of global land cover data. The first workshop was hosted by the Institute of Environmental Sustainability of the European Commission's Joint Research Center, Ispra, Italy, in March, 2003, and the second by the Boston University Center for Remote Sensing, in February, 2004. At these workshops, working group members discussed current and desired future practices for validation of global land cover maps based on remotely-sensed data. As a consensus emerged, the group developed an outline for a document that would summarize the issues involved in global land cover validation and identify recommended approaches and techniques. Following the workshops, the final document was written in sections and edited for uniformity.

1.1. Global Land Cover from Space

Within the last few years, large volumes of high-quality global remotely-sensed data have become available, provided by such orbiting instruments as SPOT-Vegetation (CNES, 2000), MODIS (Justice *et al.*, 1998), and MERIS (ESA, 2004). These imagers provide near-daily multispectral imaging of the Earth's land surface at resolutions ranging from 250 to 1000 m. Their frequent coverage provides a higher probability of observing the surface without interference from clouds, thus allowing the construction of global datasets in which nearly all points on the Earth's land surface have been imaged on multiple occasions. This, in turn, opens the door for global science data products derived from multispectral and multitemporal measurements.

Among these science data products is global land cover, typically presented as a digital thematic map in raster format with pixels in the range of 500-1000 m. Thus far, global land cover maps have been constructed using data from AVHRR (Loveland *et al.*, 2000), SPOT-Vegetation (*e.g.*, Bartalev *et al.*, 2003), and MODIS (Friedl *et al.*, 2002), and future maps are also planned from MERIS (GLOBCOVER project). Remotely-sensed global land cover products typically recognize

a limited set of land cover types, based on both multispectral signals and the change in those multispectral signals through an annual cycle. The result is normally a map with a legend that distinguishes among land covers based on vegetation form and cover – for example, deciduous and evergreen forests, woodlands, savannas, or shrublands. Nonvegetated surfaces, such as barren ground and snow or ice, are also distinguished by the spectral and temporal signal. Agriculture is typically included, but since human activity cannot be sensed directly, some types of agriculture may be omitted (*e.g.*, pastures) or recognized only with some difficulty.

Land cover at the global scale is highly useful information and has already found wide use within the scientific community (*e.g.*, Schneider *et al.*, 2003; Gerten *et al.*, 2004; Tian *et al.*, 2004; Zhou *et al.*, 2003; Gao *et al.*, 2003; Myhre *et al.*, 2003). Of primary interest is the use of land cover type to parameterize global- and continental-scale models, for example climate or carbon models. In these cases, the land covers are assigned physical attributes, such as roughness length, surface resistance to evapotranspiration, albedo, or photosynthetic efficiency (*e.g.*, Sellers *et al.*, 1994). In others, land cover may be used as an index to guide an algorithm producing another type of science data product (Lotsch *et al.*, 2003). Land cover at the global or continental scale can also be used for some aspects of land management, such as sensing regional patterns of habitat or identifying large areas suitable for conservation management (Muchoney and Strahler, 2002).

In many applications, remotely-sensed global land cover maps are simply ingested without concern for their quality or accuracy. The rationale for this action is often that conventional sources of land cover information are so generalized that anything is an improvement. Another factor leading to unquestioned use is that other uncertainties may have a greater effect on the modeled outcome than errors in land cover information. In either case, land cover maps are being used without an appreciation of their inherent uncertainties, which may be large. It is clear that users of land cover information can improve their products and predictions by having some knowledge of the error structure of the land cover data in use. Moreover, global land cover maps differ significantly, depending on the quality of the input data and the classification algorithm used to produce them, as well as the spatial resolution and legend (Townshend, *et al.*, 1991). Given this variation, the choice of a particular map may substantially affect user's outputs.

While only a few global land cover maps have been produced thus far, we may expect more to appear in the future. For example, the MODIS land cover team is producing annual versions of its global map, beginning with 2000 through the life of the MODIS mission. The intent is not to document interannual change, but rather to provide the best possible map using data from a particular year. The GLC2000 effort, which used 1-km SPOT-Vegetation data from 2000, will be repeated with finer-resolution data from the MERIS instrument acquired in 2005. Global land cover is a science data product that will be produced by the future NPOESS system on a quarterly basis (Townshend and Justice, 2002). In addition to these global efforts, regional- and continental-scale efforts such as Africover (FAO, 2004), CORINE in Europe (EEA, 1995), and MRLC2001 in the United States (USGS-EDC, 2003) approach global size and scale and thus may benefit from the perspectives on validation that we provide in this document.

1.2. Challenges to Validation

The purpose of this report is to identify useful and desirable methods and approaches to the validation of global land cover maps.[9] Here, we define the term *validation* as a suite of techniques for determining the quality of a particular map. The techniques include assessing the accuracy of a given map based on observations such as overall accuracy, errors of omission and commission by land cover class, errors analyzed by region, and fuzzy accuracy (probability of class membership), all of which may be estimated by statistical sampling. Although the validation techniques we will describe rely heavily on probability sampling designs for collecting validation data, information obtained without a proper statistical sample design will often be useful in understanding the basic

[9]Although our effort is focused on global-scale maps, many of the concepts and ideas we present at applicable at regional and local scales.

error structure of the map. Such information includes spatially-distributed confidence values provided by classification algorithms, as well as systematic qualitative examinations of the map and comparisons (both qualitative and quantitative) with other maps and data sources.

Global, coarse-resolution land cover maps constructed from remotely-sensed data are limited in the accuracies they can achieve. Before discussing validation further, it will be useful to review some of the inherent challenges to making global land cover maps and assessing their accuracies. The first of these involve limitations imposed by the satellite sensor data themselves: spectral data quality and geolocation.

- *Spectral Data Quality*. Radiometric observations of the land surface from satellite are subject to many influences beyond those of interaction with the surface (which produces the signal of interest). Instrument effects, such as detector calibration, must be carefully and consistently removed. Atmospheric effects are more problematic; these include both wavelength-dependent radiative transfer and contamination by clouds. Angular effects cause the radiance of the surface to vary with both look and illumination angle in a complex way. However, the current generation of satellite instruments yield data (typically as surface reflectances) that are corrected for such effects.
- *Geolocation*. Because the temporal change in the signal carries much of the information that discriminates among land cover types, it is important to observe each point on the ground consistently over the annual time period. Thus, the ground location associated with each pixel in each of the multitemporal images must be known to high accuracy. This, in turn, means knowing spacecraft position and velocity, as well as various instrument imaging parameters, to high precision. Topographic height of the terrain must also be known. Geolocation can be less accurate in homogeneous regions, but in areas of fine spatial pattern, achieving accurate geolocation can be demanding. Note also that geolocation accuracy is often variable within a single image or scan, because view angle will amplify geolocation errors. Overlaying images from different dates can also involve resampling, which is a process that can introduce errors of several types. Note that for some studies, absolute geolocation is not necessary, since multitemporal images can be co-registered. However, these will be limited in number and geographic scope and subject to additional errors, such as mislocation of control points and often topographic displacement that varies with view angle.

Beyond concerns associated with instruments, spacecrafts, and orbits, there are challenges associated with the land cover map as an abstraction of the nature of the land surface at a given point in space and time. Among these are legend definitions and mixed pixels.

- *Legends*. Land cover legends, for a variety of reasons, are not always comprised of exhaustive, mutually exclusive classes. For example, wetlands may be a desired class, but a wetlands pixel might also belong to forest or grassland classes. Clear rules are needed to deal with such equivocal cases. The problem of legend classes that have an anthropogenic component that is not directly remotely sensible, such as agriculture, has already been mentioned.
- *Mixed pixels*. Given the large size of the field of view (FOV) of global-scale imagers, the radiometric response of a single measurement is often generated by more than one land cover type. This phenomenon raises the issue of assigning proper and consistent land cover type labels when pixels are mixed or vary continuously on a spatial gradient. This problem is obviously exacerbated by geolocation errors.

Given a sequence of registered multispectral and multitemporal images, a classification process is used to assign a land cover type label to each pixel. The most successful classification procedures are empirical in nature, typically functioning by comparing the vector of pixel-based observations to a database of examples of such observations drawn from the land cover types identified in the legend. The examples are referred to as the *training data*. Challenges to the classification process

include an inherently high signal variance coupled with the difficulty of obtaining consistent and accurate training data.

- *High signal variance.* Most land cover types at a global scale include a wide range of variation in vegetation cover, plant structure, and understory or background condition, as well as significant variance in the way the vegetation cover changes through the annual cycle. In such a high-variance environment, classification algorithms often need considerable tuning to optimize their accuracy. At times, ancillary data, such as coarse-scale maps of agriculture or even nighttime images of city lights, may also be needed to achieve acceptable accuracies for specific classes.
- *Obtaining global training data.* Acquiring accurate and consistent training data at a global scale requires substantial effort. Typically, the only practical way to identify training sites is to use fine-resolution imagery, such as SPOT-HRV or Landsat ETM+, from one or two dates, and then determine land cover type by photointerpretation (possibly assisted by image processing techniques). Since this is not an exact science, there will be an inherent level of error in the training site database. Note also that large numbers of training sites may be needed to cover the full range of multispectral and multitemporal variance in each broad land cover class.
- *Registration and temporal change.* Once obtained, training data must be registered to the coarse-resolution data. This requires accurate geolocation of the fine-resolution imagery, which can be problematic in some situations. Also, training data may be acquired from fine-resolution images that are not necessarily contemporaneous with the coarse-resolution data. Both of these factors add to errors in the training data.

Determining the accuracy of a global land cover map also poses major challenges that will be discussed at length later in this document. Challenges to accuracy assessment include:

- *Accuracy parameters and definitions.* Accuracy of a thematic map can be defined in many ways. Overall accuracy is obviously a useful parameter, but it clearly does not tell the whole story. Per-class accuracies provide more specificity and indicate which classes are easy to map and which are harder. However, they can be defined in two different ways, taking into account user's or producer's viewpoints. Accuracy implies a comparison between the map and reference information, and this comparison requires rules to carry it out consistently. In some cases, the comparison may have more than a binary outcome. In fuzzy matching, for example, some mismatches are more acceptable and less "wrong" than others.
- *Sample design.* Accuracy can be rigorously assessed using a statistically valid sampling design. An efficient design will typically combine random stratified sampling with cluster sampling, and thus require careful planning and execution. As we will note in following sections, many variations on this theme are possible, depending on both the objectives of the accuracy study and the resources available.
- *Global sampling.* If accuracy is to be determined from a probability sample, all parts of the map must be available to the sampling process. For a global map, this implies obtaining reference land cover information at any location on land, a requirement that can only be met in a practical way by using fine-resolution remote sensing. As noted in the training site discussion above, determination of land cover using fine-resolution imagery is not without its own error. Challenges of global sampling include cost and availability of data.

1.3. Validation as a Process

Validation may have several components. These include the following, which are developed in more depth in following sections of this document.

- *Statistical observations*. Of primary importance are observations of accuracy with statistical merit – that is, unbiased estimates of accuracy measures and the variance of these estimates that are obtained by probability sampling. These are probably the most generally useful parameters, since they allow a user not only to weigh the magnitude of an error, but also to estimate the impact of that error on a model or other process using the land cover information as an input. A statistically valid design for estimating accuracy parameters has three parts. The response design specifies which data are to be collected at each sample location; the sampling design specifies the locations at which the response data are to be acquired; and the analysis lays out the formulas and tests to be applied to the observations. These parts are discussed more fully in Section 3. Our report concludes that all validation efforts should include proper estimation of accuracy parameters using a probability-based sample design, even though costs may be significant.
- *Confidence maps*. A classification algorithm will often provide a measure of confidence that quantifies how closely a classified observation matches the exemplars of the training set. Although not necessarily related to accuracy, such a confidence measure will tend to follow true accuracy if the training set is extensive and well-selected (McIver and Friedl, 2001). Such confidence measures are available for each pixel and can thus be displayed as a map. Confidence measures are discussed further in Section 2.6.
- *Other comparisons*. Comparisons between the target land cover map and other sources of land cover data can also be useful. These comparisons will not necessarily be based on data collected using a probability sample. For example, a low-cost method for assessing overall and per-class accuracy is to withhold a sample of training observations from the classification process and then use those observations as test data. While the outcome is not free of bias (*e.g.*, if the training data were collected only from areas of homogeneous land cover), it can indicate the relative magnitude of the different kinds of errors likely to be found in the map. Also useful are comparisons with other existing datasets of comparable scales. For example, a land cover map of a single continent might be overlain on one continent of a global map and disagreements tabulated. Although legend incompatibilities can be a problem, such comparisons can identify areas of disagreement that may need more work for resolution.
- *Qualitative-systematic accuracy reviews*. Another useful approach to accuracy assessment is the systematic review of the global land cover map, referred to in this document as systematic quality control. In this process, the map is divided into regular subregions, for example, on a latitude-longitude grid, and each subregion is examined separately to determine its accuracy. Examination is typically qualitative, using existing map sources, imagery, and expert knowledge to assess the map within the subregion. If carried out before the final stage of map preparation, this exercise can identify regions where classification and label assignments can be improved. Qualitative-systematic accuracy review is discussed further in Section 4.
- *Validation of land cover change*. Validation of land cover change presents its own unique set of problems. It is easy to validate errors of commission by examining pixels that are identified as having changed, but because change is relatively rare, it is hard to validate errors of omission among large numbers of pixels that are identified as unchanged. Because change is associated with a particular time interval, change training sets cannot be reused. If change is to be determined by overlaying successive thematic maps, misclassifications in either map will spuriously appear as change. Validation of land cover change is discussed in Section 5.

The components identified above all contribute toward a convergence of evidence on the validation of a global land cover product. They allow users to construct error analyses that assess how the weaknesses and strengths of a specific land cover product used as an information source affect their work. An underlying construct of the approach to validation is the dependence on design-based statistical inference to provide scientific credibility to the assessment. A practical

problem associated with this approach is the high cost of carrying out a global probabilistic sampling design, both in the effort required to collect and analyze a sufficient sample and in acquisition of the fine-resolution imagery that makes it possible.

We urge map producers, as well as funding agencies, to accept the challenge of providing proper, statistically-based accuracy assessments. A validation plan and sample design should be part of every proposed and funded effort to map global land cover. As a guideline, producing a global land cover map should consist of three more-or-less equal parts: data preparation, classification, and validation. Without proper validation, any land cover map, whether at global, regional, or local scale, remains an untested hypothesis. This document summarizes the state of the art of best practices for such validation.

2. Accuracy Assessment

2.1. Introduction

The main objective of accuracy assessment is to derive a quantitative description of the accuracy of the global land cover map. This is a nontrivial task, and it must recognized that there is no one universal "best" method of accuracy assessment, but rather a suite of methods of varying value and applicability for any given map and purpose. The selection of an approach for map accuracy assessment should recognize both the limits of the data (*e.g.,* impacts of mixed pixels) and purpose of the accuracy assessment (*e.g.,* the different accuracy requirements of diverse user communities or the needs of map producers in evaluating mapping methods *etc.*). Map accuracy assessment is very much a topic for research. Much of what follows is suggested guidelines for general use based on current practices that are commonly used in remote sensing and with which there is some familiarity among the research and user communities. Thus, the focus is on standard assessments made on a per-pixel basis, although much of the discussion is applicable to analyses undertaken on a different basis.

In accuracy assessment, we assume the following priorities for specific accuracy measurements:

1. An overall measure of map accuracy – that is, a single statement to provide an index of the general quality of the thematic map. As this is an estimate of the overall accuracy of the map it should be accompanied by confidence limits.
2. Recognizing that many users will be interested in a specific class or subset of the classes depicted on the map, measures of accuracy on a per-class basis are desired.
3. Recognizing that the overall measure of accuracy is a global statistic and that accuracy may vary locally within the map, some measure(s) of spatial variation in accuracy or related variables (*e.g.,* allocation uncertainty) should be provided. Accuracy could, for example, be calculated for defined regions (*e.g.,* continents, countries) or uncertainty metrics calculated for every pixel to indicate the confidence in the class label allocated to the pixel (see Section 2.6).

2.2. Issues and Constraints of Concern

There are many issues to be considered in an accuracy assessment (*e.g.,* Congalton and Green, 1999; Foody, 2002), but the following are of particular concern. Some of these have been introduced already in Section 1 and some are discussed further in following sections.

1. It is effectively impossible to produce a land cover map that is completely accurate and satisfies the needs of all (Brown *et al.,* 1999). The different viewpoints and components of classification accuracy also act to ensure that there is no single all-purpose universal measure of accuracy. The purpose of the map should, therefore, be considered in its production and assessment. In most mapping applications and map evaluations, interest is focused on overall map accuracy. It may, however, be more appropriate in some circumstances to focus on other features (Lark, 1995; Boschetti *et al.,* 2004). This has important implications to the evaluation of map accuracy. Commonly, a relatively subjectively defined target of greater than 85 percent overall accuracy with reasonably equal accuracy across the classes is specified, but this need not be appropriate for all maps or applications.
2. To avoid bias, a sample of pixels independent of that used to train a classification should be used in the accuracy assessment (Swain, 1978; Hammond and Verbyla, 1996). The sample design used to acquire the testing set of samples used to evaluate classification accuracy is of fundamental importance and must be considered when undertaking an accuracy assessment and interpreting the accuracy metrics derived (Stehman and Czaplewski, 1998; Stehman, 1995, 1999*a*). Sampling strategies are discussed more fully in Section 3.

3. Since the accuracy assessment is based on a sample of cases, confidence intervals should ideally accompany the metrics of accuracy contained in an accuracy statement (Rosenfield *et al.*, 1982; Thomas and Allcock, 1984).

4. The nature of the techniques used to map land cover from the remotely sensed imagery has important implications. For example, with some classifiers it is relatively easy to derive a measure of the uncertainty of the class allocation made for each pixel (*e.g.*, maximum likelihood classification), while with others the ability to derive an uncertainty metric is limited (*e.g.*, parallelepiped classification).

5. The use of site-specific approaches to accuracy assessment based on the confusion matrix requires accurate registration of the map and ground data sets. Some degree of tolerance to mislocation can be integrated into accuracy assessment (Hagen, 2003), although most assessments assume implicitly that the data sets are perfectly registered. The importance of misregistration as a source of nonthematic error in the confusion matrix is most apparent in regions where the land cover mosaic is fragmented (Estes *et al.*, 1999; Loveland *et al.*, 1999).

6. For conventional (hard) classifications, in which each image pixel is allocated to a single class, it is assumed that the pixels are pure (*i.e.*, each pixel represents an area that comprises homogeneous cover of a single land cover class). Any hard class allocation made for a mixed pixel will, to some extent, be erroneous, and alternative approaches to accuracy assessment (*e.g.*, Gopal and Woodcock, 1994; Foody, 1996; Shalan *et al.*, 2004) should be adopted if the proportion of mixed pixels is large. In general, the proportion of mixed pixels increases with a coarsening of the spatial resolution of the imagery.

7. Errors are commonly treated as being of equal magnitude. If some errors are more damaging than others, it may be possible to weight their effect in the assessment of classification accuracy (*e.g.*, Foody *et al.*, 1996; Naesset, 1996*a*; Stehman, 1999*b*; Smits *et al.*, 1999).

8. The ground or reference data may contain error and thus misclassification does not always indicate a mistake in the classification used to derive the map. In reality, therefore, the assessment of maps commonly undertaken is one of agreement or correspondence with the ground data rather than strictly of thematic accuracy. In some instances, it may be useful to include some measure of confidence in the ground data used (Scepan, 1999; Estes *et al.*, 1999).

9. The pixel is the basic spatial unit of the analysis. Maps could be produced using other spatial units. For example, the minimum mapping unit could be set at a size larger than the image pixel size. The use of large units may help in reducing the effect of spatial misregistration problems. With soft/fuzzy classifications and with super-resolution mapping, where the aim is to map at a scale finer than the source data, the problems of spatial misregistration in conventional approaches to accuracy assessment are likely to be large.

10. The same set of class definitions/protocols should be used in the image classification as in the ground data; that is, the class labels used in both data sets should have the same meaning. Approaches to explore and accommodate differences in the meaning of class labels may be useful if the classes have been defined differently in the data sets (Comber *et al.*, 2004). If different classification schemes have been used, it is still possible to evaluate the level of agreement between a map and the ground data using a cross-tabulation of class labels (*e.g.*, Finn, 1993).

11. The confusion matrix should be presented as well as the summary metrics of accuracy derived from it. To avoid problems associated with normalization (Stehman, 2004*a*), the raw matrix should be provided and the sample design used in its generation specified.

2.3. Basic Approach

The basis of the suggested approach to accuracy assessment is the confusion or error matrix. This matrix provides a cross tabulation of the class label predicted by the image classification analysis against that observed in the ground data for the test sites (Figure 2.1). The confusion matrix provides a great wealth of information on a classification. It may, for example, be used to provide overall and per-class summary metrics of land cover classification accuracy (Congalton, 1991; Congalton and Green, 1999; Foody, 2002) as well as to refine areal estimates (*e.g.*, Prisley and Smith, 1987; Hay, 1988; Jupp, 1989) or aspects of the classification analysis in order to meet specific user requirements (Lark, 1995; Smits *et al.*, 1999). Moreover, the confusion matrix is relatively easy to interpret and is familiar to both the map user and producer communities.

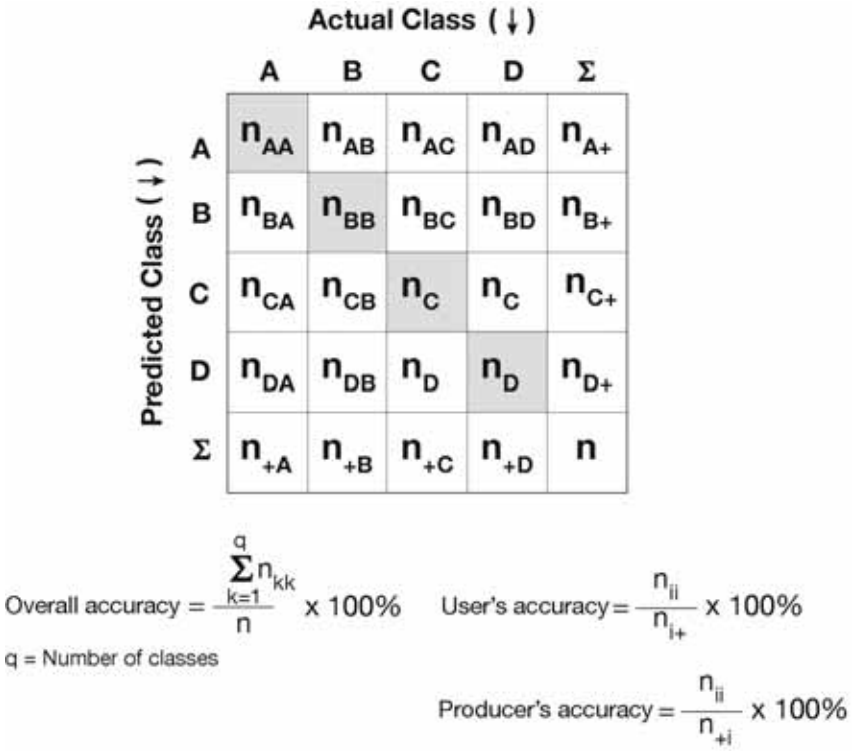


Figure 2.1. Layout of a typical confusion or error matrix, showing computation of user's and producer's accuracies.

The use of the confusion matrix in accuracy assessment applications is based on a number of important assumptions. In particular, it is assumed that each pixel can be allocated to a single class in both the ground and map data sets, and that these two data sets have the same spatial resolution and are perfectly registered. All of these assumptions are often not satisfied in remote sensing. In some instances, deviation from the assumed condition is relatively unimportant (*e.g.*, if testing pixels are drawn from very large homogenous regions of the classes then the impact of misregistration of the data sets is unlikely to have a major impact on accuracy assessment) but in other situations they may lead to significant error and misinterpretation (*e.g.*, if the land cover mosaic is very fragmented and mixed pixels are common).

Interpretation of the confusion matrix also requires consideration of the sample design used to acquire the testing set. Since the testing set is a sample, its relationship to the population (the map) is important. Confusion matrices and associated metrics of accuracy derived from a land cover map using simple random or stratified random sampling may, for example, differ markedly if there are interclass differences in the accuracy of classification. Ideally a probability sample design should be used (Stehman, 1999*a*) and this is discussed further in Section 3.

Map accuracy may be assessed using a variety of units (*e.g.,* pixels, blocks of pixels or polygons such as land parcels). For the purposes of this report it is assumed that the accuracy assessment is based on pixels. Given that the pixel is the smallest spatial unit, assessing map accuracy on a per-pixel basis is somewhat ambitious. A coarser minimum mapping unit may be more appropriate, but pixel-based assessment is common and, providing its limitations are realized, can be useful. Given that there is a trade-off between accuracy and spatial resolution, with aggregation acting to reduce misregistration errors, knowledge of the relationship between accuracy and resolution may help in the specification of an appropriate cell size for a map (Carmel, 2004).

2.4. Thematic Accuracy (Hard Classification)

For global land cover maps, accuracy assessment aims to provide an index of how closely the derived class allocations depicted in the thematic land cover map represent reality. In essence, the summary metrics of accuracy provide a measure of the degree of correctness in the class allocations in the map. Attention is, therefore, focused on thematic accuracy. The confusion matrix is well suited to this task (Figure 2.1). The cases that lie on the main diagonal of the matrix represent those correctly allocated, while those in the off-diagonal elements represent errors. Two types of thematic error, omission and commission, are possible and both may be readily derived from a confusion matrix (Congalton and Green, 1999). An error of omission occurs when a case belonging to a class is not allocated to that class by the classification. Such a case has been erroneously allocated to another class, which suffers an error of commission.

A major problem in the use of the confusion matrix and associated accuracy metrics, however, is that it may contain nonthematic error. In particular, error due to misregistration of the data sets is commonly included (Canters, 1997; Pontius, 2000; Powell *et al.,* 2004). It is important to be aware of this source of error, as the error due to misregistration may be larger than the thematic error actually present in the map. Sometimes it may be appropriate to spatially adjust locations of testing sites to account for known misregistration effects (Husak *et al.,* 1999) or to attempt to directly include some tolerance to spatial misregistration effects into the accuracy assessment (Hagen, 2003).

2.4.1. Measures of Accuracy

A variety of measures of overall and per-class accuracy can be derived from the confusion matrix. Throughout the discussion that follows, it is important to note that since the ground data are themselves a classification that may contain error it is agreement with the ground data rather than accuracy that is actually assessed.

Metrics of overall accuracy provide an indication of the quality of the entire land cover map. For overall accuracy, attention is focused on the main diagonal of the confusion matrix. Many summary metrics may be derived from a confusion matrix to express accuracy. The two most widely used measures of land cover map accuracy are the percentage of correctly allocated cases and the kappa coefficient of agreement (Trodd, 1995). These give a guide to the overall quality of the map. Although the kappa coefficient has been widely promoted for accuracy assessment (*e.g.,* Congalton *et al.,* 1983; Smits *et al.,* 1999), there are sufficient concerns with its use (*e.g.,* Foody, 1992; Ma and Redmond, 1995; Stehman and Czaplewski, 1998; Turk, 2002) that it cannot be recommended as general measure of map accuracy.

Sometimes interest is focused on the accuracy with which a particular land cover class is represented. Metrics to describe per-class accuracy can be readily derived from the confusion matrix. Clearly, this may be approached from two perspectives, depending on whether the data in the confusion matrix are read vertically or horizontally (Story and Congalton, 1986). If attention is focused on the accuracy of the map as a predictive device, concern is with errors of commission. In this situation what is generally termed user's accuracy may be derived, which is based on the ratio of correctly allocated cases of a class relative to the total number of testing cases allocated to that class. The resulting metric provides an indication of the probability that a pixel allocated to a particular land cover class actually represents that class on the ground. Reading the matrix in the

alternative way, from the map producer's perspective, the focus is on errors of omission. What is generally termed producer's accuracy may be derived from the ratio of cases correctly allocated to a class to the total number of cases of that class in the testing set. User's and producer's accuracy, therefore, convey different information and since one may be traded for the other (Lark, 1995; Boschetti *et al.*, 2004), it is important that the measure appropriate for the task in-hand is used.

Other metrics of overall and per-class accuracy can be derived from a confusion matrix (*e.g.*, Foody, 1992; Finn, 1993; Ma and Redmond, 1995; Naesset, 1996*b*; Stehman, 1997*a*). Each metric focuses on different aspects of accuracy and may vary in utility between map users. Since it is impossible to anticipate the needs of all users, the confusion matrix itself should be provided so that the user may derive a specific measure of interest. To maintain flexibility, the raw and not a normalized matrix should be provided (Stehman, 2004*a*).

Commonly in accuracy assessment, errors are treated as if being of equal magnitude. Often, however, errors vary in importance and there may be a desire to accommodate for differences in error severity in the accuracy assessment. For example, errors are often between relatively similar classes lying on either side of arbitrary class boundaries fitted to continua (Campbell and Mortenson, 1989; Sheppard *et al.*, 1995; Foody, 2000*a*). Errors can, therefore, vary from being relatively minor and insignificant to very damaging, depending on user needs. It is possible to weight errors in accuracy assessment. For example, a weighted kappa coefficient can be derived if the relative magnitude of the possible errors can be quantified (Foody *et al.*, 1996; Naesset, 1996*a*). Although this type of approach allows differences in error magnitude to be accommodated in the accuracy assessment, a major concern is the subjective nature of the weighting scheme.

Sometimes users may wish to collapse classes depicted on the land cover map. This happens typically when concern is focused on a particular broad category. For example, when interested in monitoring deforestation, a user may be willing to aggregate all different forest type classes depicted on a map into a single class. Collapsing classes may be achieved by simply aggregating the cases of the relevant classes and relabeling them as appropriate. In terms of accuracy assessment, this collapsing of classes results in the production of a new, smaller, confusion matrix and generally has the effect of increasing accuracy, as much of the error that occurred with the original set of class labels was between the classes aggregated (Foody and Embashi, 1995; DeFries and Los, 1999).

Use of a hierarchical classification scheme can facilitate collapsing classes. Thus, for example, the widely used Anderson scheme has four levels (Lillesand *et al.*, 2003). At the top level there are broad land cover classes which may be progressively broken down into more detailed classes at lower levels of the hierarchy. As an example, the class forest could be defined at a high level, at the level below this could lie the classes of deciduous, coniferous, and mixed forests, and beneath that level could be a set of classes comprising individual tree species assemblages. Reflecting the increasing precision in class definition, the accuracy of classification typically declines with progression down from the top of the hierarchy.

2.4.2. Spatial Variation in Accuracy

Conventional methods of accuracy assessment, whether of overall accuracy or on a per-class basis, are "global," in that they provide a single summary metric of the quality of the entire map. Accuracy may, however, vary within the map and some users may only be interested in parts of the mapped area. Thus many users, especially those using the map within spatially distributed models, may benefit from a spatial representation of map quality. To indicate the spatial variation in map accuracy it may be possible to derive a local estimate of map accuracy (Foody, 2005) or use a measure of the uncertainty associated with per-pixel class allocations as a guide to map quality and its spatial variation (Corves and Place, 1994; Maselli *et al.*, 1994; Foody, 2000*b*; McIver and Friedl, 2001).

One concern with the use of such measures of the strength of class membership is to note that the distinction between relative and absolute class membership is important. Here, relative

membership refers to a measure of similarity calculated with respect to exemplars of all classes, whereas absolute membership refers to a similarity measure calculated with respect to only one class. With a relative measure of class membership, such as likelihoods or posterior probabilities, there is a danger of confident misallocation. That is, an observation may fit one class somewhat poorly while fitting the remainder very poorly, yielding a high likelihood or posterior probability. In such a case, an absolute measure of class membership, such as typicality (which is related to Mahalanobis distance), may be more appropriate (Foody, 2000*b*). Often the provision of both relative and absolute measures of membership and other indicators of classification uncertainty (*e.g.,* entropy) would be useful for some users. Since such measures may be derived as a by-product of some commonly used classification algorithms, their provision to users does not place major demands on map producers, although users may need training in their interpretation.

2.5. Alternative Approaches (Soft and Fuzzy Classification)

As noted above, conventional accuracy assessment is based on the notion that the field to be mapped can be divided unambiguously into categories or themes. Additionally, it is assumed that each pixel in an image can be correctly allocated to a single theme. In essence, this model for thematic maps is based on crisp set theory, in which the legend consists of an exhaustive set of mutually exclusive classes. These simplifying assumptions help make much of the rigorous statistical analysis described above possible. However, the crisp model works better in some cases than others. As previously noted, the problem of mixed pixels is serious for land cover classification at coarse resolutions, and hence any assignment to a single class must to some extent be erroneous.

However, problems with the use of crisp sets extend beyond the problems of mixed pixels. Given the discrete nature of thematic classes, some observed land covers are not an ideal fit for any class or are suitably described by more than one class label. This problem is exacerbated when (1) the legend for the map is incomplete or does not account for all possible land covers, or (2) the land cover actually observed could fit into more than one class in the legend.

One undesirable result of the use of crisp sets is that all wrong answers are treated as completely and equally wrong. In reality, some errors in land cover maps are worse than others. For example, confusing different kinds of agricultural classes is probably a lesser problem for some users than confusing an agricultural class with an urban class. Also, mixed pixels with substantial components, but not the dominant component, of the class in the map are normally considered to be labeled entirely wrong.

There are two approaches available for the problems associated with the use of crisp set theory as outlined above. One approach is “soft” classification, which allows for estimation of the fractions of thematic classes in a single pixel. This approach retains the assumption that each place on the ground can unambiguously be assigned to a single theme in the map, but it changes the spatial scale at which the themes are manifest. A second approach is to explicitly acknowledge the possible uncertainty about the membership of accuracy assessment samples using fuzzy set theory to characterize the degree of membership of a sample in each possible thematic class (Woodcock and Gopal, 2000; Foody, 2002).

There are now a number of ways of generating soft classification maps, which show subpixel fractions of classes. Initially, the most common approach was to use linear mixture models, but there are now ways based on techniques such as maximum likelihood classification, neural networks, or decision trees (Atkinson *et al.*, 1997; Ju *et al.*, 2003; Liu and Wu, 2005). Most applications of “soft” classification retain the underlying assumption of crisp sets by constraining the subpixel proportions of the various classes to sum to unity. The assessment of the accuracy of these proportions of classes within pixels requires substantially different methods than for a thematic map in which each location is assigned to a single class. As a result, the assessment of the accuracy of soft classifications is not treated here in great detail. The most common measure of the accuracy of such estimates is the root-mean-square-error (RMSE), which can be calculated for all k classes in a soft classification as follows:

$$RMSE = \left[\frac{1}{nk} \sum_{i=1}^{n} \sum_{j=1}^{k} \left(x_{ij} - y_{ij} \right)^2 \right]^{\frac{1}{2}} \quad (2.1)$$

where there are *n* samples of *k* classes, and *x* is the map estimate and *y* the reference measurement of the proportions. One of the benefits of the use of RMSE as a measure of accuracy is that it is in the units of the variable in question (subpixel proportions in this case). It is worth noting that all the discussion presented above regarding the selection of sample sites for accuracy assessment (Section 2.3) is also relevant here.

Fuzzy sets present an alternative to crisp sets in accuracy assessment of thematic maps. Instead of selecting a single class that is assumed to be completely correct and all others completely wrong, it is possible to provide levels of membership for the sample site in each thematic class. In this formulation a sample site x has a membership value μ that ranges between 0 and 1 for the class γ that has been assigned to the sample site in the map. Additionally, a membership value is assigned for all classes in the map, denoted as the set of classes Γ. In the formulation below, map classes other than the one assigned to a sample site use the notation of γ'. Using data of this kind, it is possible to provide some additional measures of map accuracy above and beyond those described above for crisp sets. For example, Gopal and Woodcock (1994) provide metrics of MAX, RIGHT, and DIFFERENCE. The MAX operator is the same as conventional overall accuracy, and can be calculated as

$$\text{MAX}(x,\gamma) = \begin{cases} 1 & \text{if } \mu_\gamma(x) \geq \mu_{\gamma'}(x) \text{ for all } \gamma' \in \Gamma \\ 0 & \text{Otherwise} \end{cases} \quad (2.2)$$

The percentage of sample sites for which MAX = 1 provides an extremely conservative assessment of the overall accuracy of the map. These same membership values $\mu_\gamma(x)$ can generate a conventional confusion matrix if the fuzzy membership values are forced to 1 for the class with the highest fuzzy membership score and zero for all others. Similarly, Jager and Benz (2000) provide methods for analyzing fuzzy accuracy data that include generalization to the case of crisp accuracy data.

The RIGHT operator counts all sample sites as correct that receive a membership fuzzy membership score above a certain threshold τ, and can be calculated as:

$$\text{RIGHT}(x,\gamma) = \begin{cases} 1 & \text{if } \mu_\gamma(x) \geq \tau \\ 0 & \text{Otherwise} \end{cases} \quad (2.3)$$

Note that there can be more than one map class that can be considered RIGHT for a sample site, and hence it is generally a less stringent measure of map accuracy than the MAX operator. However, it is possible for there to be one or zero classes considered RIGHT for any sample site. The RIGHT operator is an answer to the question of how frequently do users of the map find a thematic class for a site that they consider to be correct.

One interesting benefit of the use of fuzzy accuracy data is the ability to characterize the magnitude of errors (or sites not passing the stringent criteria of the MAX operator). Errors between related classes can have low DIFFERENCE values and those between unrelated classes can be high. The DIFFERENCE operator is calculated as the difference between the membership value assigned for the class observed in the map and the maximum membership value for a class assigned for a sample site:

$$\Delta(x) = \mu_\gamma(x) - \text{Max } \mu_{\gamma'}(x) \quad (2.4)$$

DIFFERENCE values near negative one represent more serious errors than those close to zero. The distribution of DIFFERENCE values for all sample sites can provide an overall indication of the magnitude of errors in the map.

Fuzzy accuracy assessment can provide additional information above and beyond that provided by conventional accuracy assessment and thus can be beneficial to certain users. However, our recommendation is that the technique be used in addition to conventional methods rather than instead of conventional methods. Note that data collected for fuzzy accuracy assessment can be easily simplified for use in conventional analyses (Jager and Benz, 2000; Woodcock and Gopal, 2000), and that sample designs appropriate to crisp classification maps can work equally well for fuzzy accuracy assessment.

2.6. Confidence-Based Quality Assessment

One limitation of conventional design-based accuracy assessment approaches is that they typically provide global information regarding overall map quality. That is, the accuracy measures apply to the entire region, but are not intended to apply to subregions within the map. Map errors are neither random nor stationary in space, so subregional estimates of accuracy are typically of interest. To obtain subregional estimates using conventional design-based accuracy assessment approaches available validation data must possess adequate sample size within the region of interest for precise estimates. Unfortunately, sufficient subregional data are rarely available to support this.

To provide more spatially explicit information on map quality, in this section we consider an approach to map accuracy assessment that is somewhat different to those that have been discussed previously in this document. We refer to this approach as "confidence-based quality assessment." This approach uses information computed by classification algorithms to provide spatially explicit representations of map quality and has recently been gaining acceptance in the remote sensing and land cover mapping community. The key difference between confidence-based quality assessment and conventional accuracy assessment methods of is that confidence-based quality assessment provides a metric of classification quality at each pixel. Thus, the user is provided with a spatially explicit representation of classification quality that supplies substantial additional information relative to conventional accuracy assessment.

Confidence-based quality assessment has two key advantages. First and most important, the user is provided with an estimate of the map quality at each pixel in the map. Second, the approach does not require additional reference data and is therefore very cost effective. Note, however, that confidence-based assessment methods are not a substitute for map accuracy assessment. Rather, they should be viewed as providing valuable and complementary information to more conventional methods.

2.6.1. Theory of Confidence-Based Quality Assessment

Confidence-based quality assessment utilizes the fact that the statistical or numerical theory underlying many classification algorithms can be used to convey information related to classification quality. Note that because the confidence measure depends solely on the classification algorithm and requires no additional data, the value of the specific metric used to quantify map quality will vary at each pixel depending on the specific classification algorithm that is used to create the map. Also, for most classification models that operate on a pixel-by-pixel basis (*i.e.*, no spatial context), the confidence measure at each pixel is generated independently from those of surrounding pixels.

A variety of different approaches to confidence-based quality assessment have been developed over the past fifteen years. These include the use of geostatistics (Kyriakidis and Dungan, 2001; de Bruin, 2000), maximum likelihood classification (Foody *et al.*, 1992), interpolation of classification errors at training sites (Steele *et al.*, 1998), and ensemble classification algorithms (McIver and Friedl, 2001; Steele *et al.*, 2003). While each approach is somewhat different, they all provide a measure of classification quality at each pixel. Below we focus on two main examples – maximum likelihood classification and classification trees.

The concept of confidence-based quality assessment was first described in detail in the remote sensing literature by Foody *et al.* (1992), who used the maximum likelihood classification

algorithm to demonstrate that statistical classification models can provide information in addition to the classification prediction at each pixel. In particular, Foody *et al.* (1992) noted that the maximum likelihood algorithm computes the *a posteriori* probability of the most likely class. In addition, the *typicality* (which is related to the Mahalanobis distance) of the pixel can also be computed. These metrics provide information regarding the statistical relationship between the vector of multispectral observations at each pixel and the training data used to estimate the classification.

The *a posteriori* probability provides information regarding the overall spectral separation among the various classes. That is, if the posterior probability associated with the most likely class is high, this suggests that the data vector associated with the pixel in question resembles the most likely class more strongly than it resembles the other classes. However, because the posterior probabilities are normalized to sum to unity, a high *a posteriori* value does not guarantee a good match with the training data. That is, if a given pixel is not similar to any of the training data, it is possible to compute a high *a posteriori* probability even though the pixel is quite different from the predicted class. To quantify the goodness of the fit, the typicality can be used to measure the distance of the observation vector from the centroid of the most likely class. Indeed, as Foody *et al.* (1992) point out, some image processing systems allow users to leave unclassified those pixels whose typicality is low, irrespective of the posterior probability.

The posterior probability and typicality are both well suited to providing confidence-based quality assessment. However, maximum likelihood classification relies on a Gaussian distance function, which assumes underlying normality. In this context, recent work by McIver and Friedl (2001) using classification trees has demonstrated that ensemble classification methods can be used to estimate measures similar to those described by Foody *et al.* (1992) while avoiding the requirement for Gaussian data. To do this, McIver and Friedl (2001) used decision trees in combination with an ensemble classification technique to classify three disparate data sets. Specifically, McIver and Friedl (2001) used a technique called *boosting*, in which multiple classifications are estimated using resampled versions of the original training data. By exploiting the work of Friedman *et al.* (2000), who showed that boosting is a form of additive logistic regression, McIver and Friedl were able to show how an approximate measure of the posterior probability could be estimated at each pixel. This measure was termed the *classification confidence*. Using cross-validation methods, they showed that both overall and class-specific classification confidence are highly correlated with the overall classification accuracy (Figure 2.2). In other words, as the classification confidence increases, the classification accuracy also tends to increase in a fairly direct fashion. This approach is used to provide a map of classification quality at each pixel for the MODIS global land cover product (Friedl *et al.*, 2002).

2.6.2. Discussion and Recommendations

In this discussion we have primarily emphasized the use of posterior probabilities to quantify classification map quality in a spatially explicit fashion. In this context, it is important to note that other metrics have also been used in this regard, such as the classification margin and entropy (Gorte, 1998). However, as we have previously indicated, the key point of this discussion is that whatever the method or metric, spatially explicit quantification of map quality at each pixel provides a valuable complement to conventional design-based assessments. To this end, several key recommendations arise from this discussion:

1. Confidence-based quality assessment provides useful information for map quality and accuracy assessment. Many classification algorithms are capable of producing estimates of a posteriori probabilities, and their use for map quality assessment should be encouraged.

2. Alternatives to a posteriori probabilities are available, and should be used in parallel with each other. In particular, confidence-based estimates of posterior probabilities depend on the nature and quality of the training data and population being mapped. Thus, care must

be taken in interpreting the results of confidence-based quality assessment, especially for those pixels exhibiting very high apparent quality.

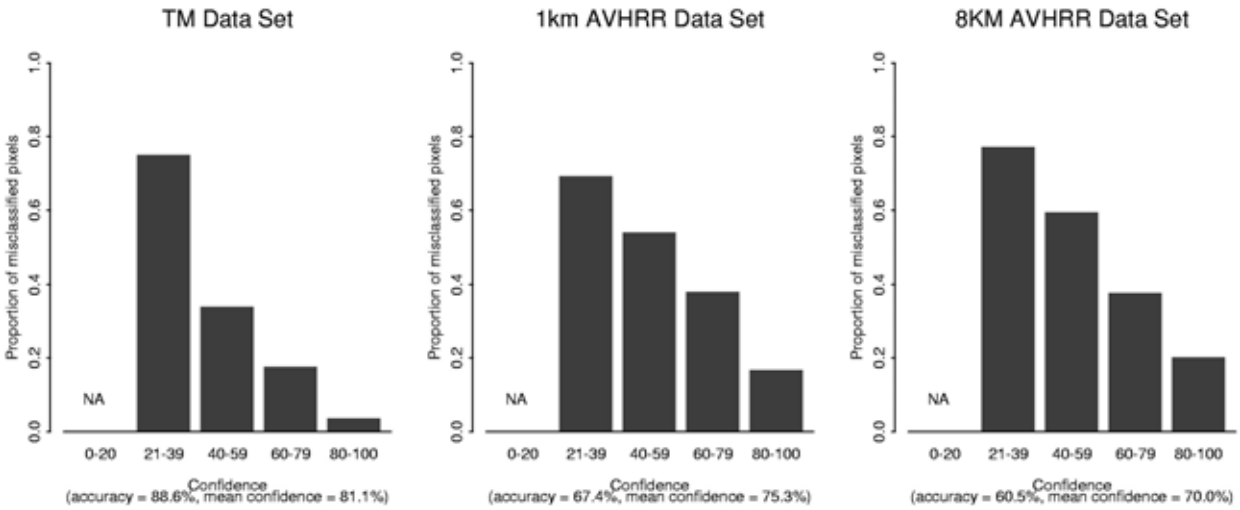


Figure 2.2. Results from McIver and Friedl (2001) showing relationship between classification confidence and classification accuracy. Barplots show cross-validated accuracy for three different data sets, binned into 20 percent ranges of confidence. In each case, the proportion of misclassified cases decreases as the classification confidence increases. See McIver and Friedl (2001) for details.

3. Strategies for Global Accuracy Assessment Using Probability Sampling

Statistically-based accuracy assessments are composed of three parts: the response design, sampling design, and analysis of the data (Stehman and Czaplewski, 1998). The response design consists of the protocols used to determine the reference or ground condition label (or labels) and the definition of agreement for comparing the map label(s) to the reference label(s). The sampling design is the protocol for selecting the locations at which the reference data are obtained. Throughout we assume that a sampling design for accuracy assessment will be implemented independently of any sampling implemented for collecting reference data to train or develop the classification. The analysis is the set of formulas for estimating the accuracy measures of interest and their associated standard errors. In this section, we will focus on the sampling strategy, which is the combination of the sampling design and the estimators.

A key feature of accuracy assessments of global land cover products is that these assessments should be statistically defensible. The requisite statistical rigor can be achieved by adhering to protocols ensuring the validity of design-based inference. "Design-based" refers to a survey sampling inference framework in which properties of estimators such as bias and variance are determined over the set of all possible samples that potentially could occur given a specified sampling design (*cf.* Särndal *et al.*, 1992). This inference framework is appropriate when the objective is to assess and describe the accuracy of a particular global land cover product. The inference framework guides the decision-making process when selecting the sampling strategy.

To satisfy requirements of design-based inference, the sampling design should be a probability sampling design, and the estimators should be constructed following the principle of consistent estimation. In addition, a third desirable design criterion is that the sampling strategy should produce accuracy estimators with adequate precision. At the planning stage, many of the decisions regarding the sampling strategy for global accuracy assessment are productively addressed by evaluating each decision in terms of these three criteria. In this section, the discussion will be directed to a pixel-based assessment, but much of the underlying theory and practical recommendations are applicable to other assessment units – for example, land cover polygons or regularly-shaped areal units such as 5 by 5 km blocks. This section focuses on how to construct the sampling design and analysis components.

Design-based inference is predicated on implementing a probability sampling design. The definition of probability sampling focuses on inclusion probabilities, where an inclusion probability is defined as the probability that a particular pixel will be chosen for the sample. Probability sampling requires these inclusion probabilities to be known for all pixels selected in the sample, and nonzero for all pixels in the population (the entire region mapped). Many probability sampling designs have been developed, including familiar designs such as simple random, systematic, stratified random, and one- and two-stage cluster sampling. Adherence to probability sampling imposes some constraints on the sampling protocol to ensure that the inclusion probabilities can be determined.

Sample selection protocols that do not qualify as probability sampling designs typically include many steps in a convoluted, ad hoc procedure, thereby making it impractical, if not impossible, to derive the inclusion probabilities. If the protocol specifies sampling only within areas of homogeneous land cover, the requirement of nonzero inclusion probabilities is not satisfied for a large portion of the map. Another nonprobabilistic sampling protocol is "balanced sampling" (Royall and Eberhardt, 1975) in which the sample is selected in a nonrandomized fashion so that specified sample characteristics match known population characteristics (*e.g.*, the sample is balanced so that the proportion of land cover in each map class matches the proportion of each class known from the full map). Although balanced sampling is motivated by a more logical and laudable rationale than most other versions of nonprobabilistic sampling, it is the lack of randomization that renders this approach inappropriate for use in a design-based inference framework. Strategies within a design-based framework exist to achieve criteria such as sample

balance, so it is unnecessary to sacrifice the rigor of design-based inference to attain desirable properties often claimed as the purview of nonprobabilistic sampling methods. But to take advantage of the powerful features of design-based inference, our recommendation is to implement a probability sampling design to provide the core data for the assessment. Nonprobabilistic sampling may be used to supplement the data from the probability sampling design.

A last motivation for implementing a probability sampling design for accuracy assessment is that the data collected can be used for other objectives that also require a statistically rigorous foundation. Specifically, accuracy assessment reference data can be used to calibrate the area estimates derived from a complete coverage map (Czaplewski and Catts, 1992; van Deusen, 1996; Gallego, 2004). The accuracy assessment reference data supply the information necessary to adjust the map-based area estimates for bias introduced by classification error. Without the underlying support of a probability sample, these calibration estimators have doubtful utility.

The second requirement of design-based inference is consistent estimation. The principle of consistent estimation requires that the inclusion probabilities are incorporated in the formulas used to estimate the accuracy metrics of interest. Consistent estimators ensure that the population parameters of interest are in fact being estimated, or more loosely phrased, that the estimators have at worst small biases. Basic estimation formulas presented in survey sampling texts (Cochran, 1977; Lohr, 1999; Thompson, 1992) are almost always consistent estimators, and these estimators are usually presented in algebraically simpler versions than those expressed directly in terms of inclusion probabilities.

If inclusion probabilities are not equal for all elements of the sample and these inclusion probabilities are ignored in the estimator, a significant bias likely results. For example, a stratified random sample with equal allocation (same sample size in each stratum) is commonly incorrectly analyzed as if the data had arisen from a simple random sample. Such an unweighted analysis fails to recognize that the inclusion probabilities are different among strata, and these unweighted estimators (not based on the inclusion probabilities) do not satisfy the consistency criterion. Specific guidelines on implementing consistent estimators are described in Section 3.3.

Implementing a probability sampling design and employing consistent estimators supplies the scientific credibility of the design-based inference approach to global map accuracy assessment. Design-based inference is a generally accepted framework for characterizing a population based on sample data. It requires minimal assumptions to justify the validity of the accuracy estimators and their associated standard errors (Stehman, 2000), and the approach is applicable to any classification scheme (*e.g.*, crisp, fuzzy, or rough) as well as to continuous fields. Other inference frameworks, for example model-based or Bayesian (Little, 2004), have received little attention in accuracy assessment (see Green and Strawderman, 1994, for an exception). This is not to suggest that these approaches to inference are not valid for accuracy assessment, but only that they have not been implemented for this use. Adopting an alternate inferential framework would likely lead to different guidelines for a global accuracy assessment sampling strategy than those recommended here.

The minimal dependence on distributional assumptions of design-based inference is an appealing feature for global accuracy assessment. An inference framework heavily dependent on a model or other assumptions would require the cumbersome task of not only explicitly identifying these assumptions and model structures, but also justifying that they were satisfied for the particular application. The multitude of uses and users of a global map would suggest that validating assumptions may be even more difficult because of the large number of different analyses to which the data would be subject. Lastly, the objectivity provided by the randomization protocol of probability sampling provides assurance that the sample has not been selected, either consciously or unconsciously, to produce favorable accuracy results.

3.1. Planning the Sampling Design

Designing an accuracy assessment requires adhering to protocols that ensure statistical rigor yet still accommodate practical realities related to cost constraints. An important first step in the planning process is to identify and prioritize the objectives of the assessment. Objectives typically include estimating overall accuracy and class specific accuracy (*e.g.*, user's and producer's accuracies), perhaps ordered by importance of the classes. Often regional accuracy estimates are desired (*e.g.*, continent, biome, ecoregion, or an administrative unit such as a state, province or country), and the objectives may extend to include estimating class-specific accuracy within each region. Global land cover change maps introduce additional objectives related to estimating accuracy of change.

The budget available to conduct the assessment and the desired precision of the estimates are additional important inputs into the planning process as these factors largely determine the sample size and allocation of sampling resources to different objectives. The final choice of sampling design typically reflects numerous compromise decisions among the many objectives, and some objectives will be satisfied better than others. For example, cost constraints may impose a smaller sample size than is adequate to achieve the target precision for all estimators specified by the objectives.

No single sampling design serves as a universally appropriate design for global assessments. However, some general recommendations for best available practice can be made. As emphasized throughout, constructing the sampling design so that it satisfies the definition of a probability sample is the most important design characteristic. This rules out those protocols for which inclusion probabilities cannot be determined, and it also eliminates from consideration the practice of sampling only from homogeneous areas of the map to diminish problems associated with inaccurate spatial co-location of the map and reference sampling units. Unfortunately, this "remedy" to the spatial registration problem violates the requirements of probability sampling, and typically leads to optimistic estimates of accuracy. Although concern with confounding thematic error with spatial registration error is justified, this issue should be addressed in the analysis (see Section 3.2).

Strata and clusters are often employed in accuracy assessment sampling designs. Strata are typically motivated by estimation objectives. For example, stratifying by map land cover class targets the objective of estimating class-specific accuracy, and stratifying by regions targets the objective of estimating region-specific accuracy. Without stratification, the sample size representing a rare class or small region may be insufficient to precisely estimate accuracy. Budget constraints often limit the number of strata that can be effectively employed. For example, stratifying by the cross-classification of both region and land cover type may be desirable. But often resources are not available to obtain a large enough sample to estimate accuracy precisely for this many strata. The practical recommendation for global accuracy assessment is to stratify to meet the highest priority objectives, but to not "over" stratify at the expense of poorer precision for other important estimates. For example, a first cut at defining strata for a global accuracy assessment may identify major regional strata (*e.g.*, continents), and then define strata for a limited number of land cover classes (*e.g.*, six to ten) within each major regional stratum. Even this moderate degree of stratification could easily produce 40-50 strata.

Stratifying by map land cover class and allocating approximately equal sample sizes to each stratum is a relatively common practice in accuracy assessment. This approach is designed to provide approximately equal precision for estimated user's accuracy of each class (assuming that all classes have approximately the same accuracy), and treats each class as equally important. Larger sample sizes can be allocated to high priority classes identified by the objectives.

It may happen that the version of the map used to create the stratification for accuracy assessment sampling is ultimately replaced, for example by an updated map reflecting improved classification rules or a map based on a modified classification scheme. The fact that the map used to construct the stratification is no longer the version of the map being assessed does not invalidate the stratified sampling design. The inclusion probabilities of the sampling design are unchanged by

revising the map because the inclusion probabilities are set in place at the time the sample is selected. Hence, the probability sampling requirement is still met. The main effect of revising the map after a stratified sample has been selected is that the originally identified strata may no longer correspond to the map classes for which class-specific accuracy is desired. The analysis can readily accommodate this change if we apply the general rules of consistent estimation (Section 3.2). In effect, the sample data are regrouped into their revised map classes, but each pixel retains its inclusion probability as determined by the original stratification.

Cluster sampling is motivated by cost. Spatially clustering the reference sample pixels lowers the overall cost of data collection, either by reducing travel time in the case of ground visits, or by reducing the total number and processing time of aerial photographs, high resolution satellite images, or videography used in the response design protocol. Clusters are particularly appropriate when the reference material establishes a natural cluster (*e.g.*, Landsat scene, aerial photograph, or videography frame). When no natural cluster is obvious, a common practice is to define the cluster as a regularly-shaped unit, for example a 5 by 5 km block. In the terminology of cluster sampling, a cluster or group of pixels is the primary sampling (PSU) and a pixel is a secondary sampling unit (SSU). The sampling design must specify how to select both PSUs and SSUs.

Both one- and two-stage cluster sampling have been employed in accuracy assessment. In one-stage cluster sampling, all SSUs within each sampled PSU are included in the sample. One-stage cluster sampling is usually only practical with relatively small PSUs, for example 5 by 5 or 3 by 3 pixel clusters. For the larger size clusters likely to be most cost-effective for a global accuracy assessment, subsampling within each cluster will be preferred because two-stage cluster sampling will be more precise than one-stage cluster sampling for the same fixed total cost. That is, if the PSUs are large, precision of the accuracy estimators will be better if more PSUs are sampled even if it means having to subsample within the selected PSUs.

The cost per pixel sampled is typically less for cluster sampling compared to other designs. Once the investment has been made to interpret or reach a single pixel within a cluster, the marginal cost of sampling additional pixels within that same cluster and determining the reference class is much lower. However, the information per pixel may be less because of spatial correlation of the response within each sample cluster. For example, pixels within a cluster may be more likely to be similar in terms of most being correctly (or incorrectly) classified relative to pixels randomly chosen from several different clusters. Cluster sampling raises the question of whether the cost saving obtained by sampling multiple pixels within each cluster translates into a large enough increase in sample size to compensate for the positive within-cluster correlation of classification error that typically occurs. Clusters also increase the complexity of standard error estimators, although this is not an insurmountable problem (Stehman, 1997*b*; Magnussen *et al.*, 2004).

Incorporating both strata and clusters may be desirable for a global accuracy assessment. Stratification by large geographic regions such as continents is likely to be desirable and class-specific accuracy is likely to be a high priority objective. Because resources for the assessment will be limited, the cost efficiency of cluster sampling is relevant. Jointly incorporating both clusters and strata in the sampling design is a trickier proposition than incorporating just one or the other (Stehman, 2004*b*).

An approach based on two-stage cluster sampling successfully incorporates both of these structures (Nusser and Klaas, 2003; Stehman *et al.*, 2003). In this design, the cluster or primary sampling unit (PSU) could be a 6 km by 6 km block, a watershed, a county or township, or a Landsat scene. A sample of PSUs is selected (the first-stage sample), and all pixels within these first-stage sample PSUs are stratified by map land cover class. A stratified random sample of pixels is selected from this list. Because only pixels within the first-stage sample PSUs are eligible to be selected by the second stage stratified design, the sample is spatially constrained so that all sample pixels are within a limited number of PSUs, thus achieving the desired spatial control over the sample. This design groups the sampled pixels into fewer PSUs than would be the case had these pixels been selected by stratified random sampling without the intervening first-stage selection of clusters (Wickham *et al.*, 2004*a*), yet the design still provides the capability to target

sampling effort to the land cover strata identified. This design has been employed in agricultural surveys (Kott, 1990), so a strong precedent exists to support its potential use for global accuracy assessment sampling.

Stratifying the PSUs themselves to direct more of the sampling resources to rare classes is an alternative possibility. A problem encountered with this alternative is that the cluster sizes that are cost effective result in most clusters containing a mixture of land cover classes. In this event, rules must be specified for assigning a cluster to a single stratum: to what land cover stratum is a PSU assigned when the PSU contains pixels of several different land cover classes? One option is to assign the PSU to the land cover stratum of the dominant land cover type in that cluster (tie-breaking rules would be needed). A problem with such a stratum assignment protocol is that if the rare class pixels are spatially dispersed, only a few PSUs may be assigned to the rare class strata. This could result in the rare class sample pixels being located in very few PSUs, which in turn would diminish precision of the accuracy estimators.

Once PSUs have been assigned to strata, a stratified sample of PSUs is selected. Because of the large size of the PSUs, one-stage cluster sampling will not be cost-effective, requiring selection of a subsample from each first-stage PSU. Employing simple random sampling as the within-PSU (second stage) design offers the advantage of simplicity, but provides no additional control over which land cover classes the sampled pixels belong to beyond what has been achieved by the stratification of the PSUs. That is, even if the rare land cover class is the most common class within the PSU, it is not guaranteed that pixels of this rare class will be selected. In exchange for a slightly simpler design protocol, this option sacrifices the stronger control of allocation of the sample to the targeted strata offered by the design stratifying the pixels within PSUs rather than stratifying the PSUs themselves.

3.2. Analysis

Given the known inclusion probabilities of a probability sampling design, consistent estimators can be constructed for most accuracy metrics used in practice, whether these metrics are derived for hard or soft classifications or for continuous fields (*e.g.*, DeFries *et al.*, 1999, 2000). The general approach will be reviewed for the standard accuracy measures applied to a hard classification scheme: overall, user's, and producer's accuracies. All of these measures can be viewed as ratios, and this provides a general framework for estimation. Suppose the numerator of the ratio is determined by condition A, and the denominator is determined by condition B, each condition defined on an individual assessment unit (*e.g.*, a pixel). For example, to represent user's accuracy of "forest," condition *A* is that a pixel mapped as forest is actually forest, and condition *B* is that the pixel is mapped as forest. User's accuracy for forest is then the total area (or number of pixels) meeting condition *A* divided by the total area (or number of pixels) meeting condition *B*. For pixel *u*, let y_u=1 if pixel *u* meets condition *A*, and y_u=0 otherwise. Similarly, let x_u=1 if pixel *u* meets condition *B*, and x_u=0 otherwise. The population parameter, in ratio form, is *Y/X*, where *Y* is the total number of pixels of condition *A* and *X* is the total number of pixels of condition *B*. A consistent estimator of this ratio is

$$\hat{R} = \hat{Y} / \hat{X} \tag{3.1}$$

where $\hat{Y} = \sum_{u \in s} y_u / \pi_u$ and $\hat{X} = \sum_{u \in s} x_u / \pi_u$ are the sample-based estimators of *Y* and *X*, π_u is the inclusion probability of pixel *u*, and summation is over all pixels in the sample. This estimator is adaptable to many different parameters. For example, to estimate an accuracy measure for a subregion of the map (*e.g.*, country), the definitions of *A* and *B* should include the condition that the sample pixel falls within that subregion. The estimator is applicable if the classification scheme is collapsed by defining conditions *A* and *B* according to the collapsed classification, or if the map is revised after initial stratification. In the latter case, the inclusion probabilities from the original selection are still appropriate.

Accuracy estimators can also be derived by first estimating the cell entries of the error matrix. Let $z_{ij,u}$=1 if pixel *u* belongs to row *i*, column *j* of the error matrix, and $z_{ij,u}$=0 otherwise. The estimated

number of pixels in cell (*i*, *j*) of the error matrix is $\sum_{u \in s} z_{ij,u} / \pi_u$ where $\sum_{u \in s}$ denotes summation over all pixels in the sample. Overall, user's and producer's accuracies are then estimated by replacing the true cell pixel counts by the estimated counts in the formulas defining each metric of accuracy. For example, user's accuracy for class *k* is estimated by $\hat{N}_{kk} / N_{k+}$ where $\hat{N}_{kk}$ is the estimated cell entry for row *k*, column *k*, and N_{k+} is the number of pixels mapped as class *k*. Producer's accuracy for class *k* would be estimated by $\hat{N}_{kk} / \hat{N}_{+k}$, where $\hat{N}_{+k} = \sum_{i=1}^{c} \hat{N}_{ik}$, *c* is the number of land cover classes, and $\hat{N}_{ik}$ is the estimated number of pixels in cell (*i*, *k*) of the error matrix. These general estimator forms can be applied to data obtained via any probability sampling design.

Reporting standard errors quantifying the variability of the accuracy estimates should be routine practice in global accuracy assessments. Ideally, these standard errors will be small, indicating that the accuracy estimates are precise. For the accuracy estimators expressed in the form of a ratio estimator, the standard error is the square root of the following general variance formula:

$$Var\left(\hat{R}\right) = \sum_{u \in s} \sum_{v \in s} \left(\frac{\pi_{uv} - \pi_u \pi_v}{\pi_{uv}} \right) \left(\frac{y_u - \hat{R} x_u}{\pi_u} \right) \left(\frac{y_v - \hat{R} x_v}{\pi_v} \right) \qquad (3.2)$$

where π_u and π_v are the inclusion probabilities for sample pixels *u* and *v*, π_{uv} is the pairwise inclusion probability for sample pixels *u* and *v* (*i.e.*, the probability that both pixels *u* and *v* would be included in the sample), and the double summation is over all pairs of pixels in the sample. Because standard error estimation is relatively complex, it is desirable to estimate these standard errors using a reliable survey sampling analysis program such as provided by SAS (Statistical Analysis Software, Cary, NC) statistical software. Otherwise, the algorithms and programs constructed to estimate standard errors will need to be carefully verified prior to accepting computations as valid.

Because many of the users of a global map will have special accuracy interests they wish to explore, it is desirable for the reference data to be available to the scientific community. In addition to including the sample locations and the reference classification, the inclusion probabilities for each sample unit will also need to be provided, accompanied by a strong recommendation to users to incorporate these inclusion probabilities in their descriptive estimates. The general formulas provided in the previous subsection supply the necessary estimation theory. It may be impractical to provide all the information necessary to generate standard errors for all anticipated user specified estimates of accuracy, particularly if the sampling design is complex (*e.g.*, multiple levels of stratification, two-stage cluster sampling). It may be possible to provide some simple variance approximations that users can apply, as for example, by generating design effects for cluster sampling (Kish, 1965). One caveat of permitting user access to the reference data is that any reference data obtained via a confidentiality agreement could obviously not be released.

Ideally, accuracy assessment ground reference data would be absolutely correct, but in practice reference data errors will be present. Congalton and Green (1993) catalog potential sources of error in reference data. Standard methods for accommodating reference data error have not been adopted either for the analysis or the format for reporting results. Although no best practices have been identified, the potential impact of reference data error should not be ignored. The best strategy is, of course, to conduct the response design protocol to reduce reference data error as much as possible. This requires adhering to carefully specified, explicit protocols for the response design, and implementing ongoing quality control checks to monitor if the data are being obtained according to these protocols. Quantifying the potential effect on accuracy of different sources of reference data error will contribute valuable information (*cf.* Powell *et al.*, 2004). For example, if multiple interpreters are providing the reference data, agreement among interpreters should be quantified by having these interpreters evaluate a common test set of reference data. As another

example, suppose that reference data are not all contemporaneous with the imagery used in the classification. Here, valid questions include what proportion of the sample is so affected, and are these temporal differences in the data sources associated with classification error (Wickham *et al.*, 2004*b*), thereby indicating a possible confounding of the time difference with true thematic error.

The potential impact on the accuracy results of spatial misregistration between the reference and map locations should also be quantified. A simple approach to this problem is to provide separate accuracy results for homogeneous areas of the map to contrast to the estimates for the full map. Location error should have a minor impact on the homogeneous subset results because the pixels neighboring the target sample pixel have the same label as the target pixel. The difference between accuracy of the homogeneous subset and the full map provides a quantitative depiction of the effect of location error on the accuracy results. Secondary analyses such as those reported in Smith *et al.* (2003) and van Oort *et al.* (2004) may be useful to establish relationships between classification error and landscape characteristics.

The presence of reference data error challenges the notion that a single definition of "agreement" is sufficient when reporting accuracy results. To gain better understanding of the potential effect of reference data error, accuracy results derived from several definitions of agreement based on different interpretations of the reference data may need to be reported. Except for the recommendation not to limit sampling to homogeneous areas, standard procedures have yet to be established for accommodating other sources of reference data error in the analysis and reporting of accuracy results.

3.3. Using Existing Data in Global Accuracy Assessments

Existing data are defined as reference data available to the accuracy assessment that would not require expending resources for field visits, imagery, or other reference data materials. Some effort may need to be invested to convert the data for use in the accuracy assessment, for example to reclassify the data to match the map classification scheme. Sources of existing data may be an ongoing environmental monitoring program, or simply archived data collected for some other purpose. Because of the high cost of collecting reference data, the question often arises whether existing data can be effectively incorporated into a global accuracy assessment.

Existing data may be the sole source of reference data, or used to supplement the reference data collected specifically to assess the global map. Typically existing data will lack the necessary coverage to serve as sole source reference data for a global map accuracy assessment, so it is this second use, supplementing the accuracy assessment sample, that is the more likely application.

Several considerations are relevant to incorporating existing data into a global accuracy assessment protocol. The existing data must first be evaluated to determine if they are compatible with the response design protocol (*e.g.*, classification scheme, spatial support of reference data) specified for the map. If this condition is satisfied, the next step would be to determine the sampling design, if any, that was used to collect these data. The ideal situation is that the existing data originated from a probability sampling design. Ongoing environmental monitoring programs such as the National Resources Inventory (Nusser and Goebel, 1997) and Forest Inventory and Analysis (USFS, 1992) in the United States are potential sources of high quality reference data originating from a probability sampling design, but problems of data confidentiality and administrative coordination may still be considerable (Stehman *et al.*, 2000). Combining two probability samples within the design-based framework can be achieved using dual frame sampling estimation methods. If the existing data have not been collected using a probability sampling design, their use may not be representative of the larger population. For example, existing data are often available because of high interest in certain locations, and these locations may have very different characteristics in terms of classification error.

Although no additional field or data acquisition costs are incurred when obtaining existing data, these data nevertheless still have costs associated with their use in a global accuracy assessment. This cost is attributable to the time expended to determine if the existing data are compatible with the reference data being collected and to develop and implement the more complex estimation

procedures. Further, the time and effort required to administer the exchange of data may be considerable, and confidentiality concerns may limit access to the data. Geographic coverage will likely be globally inconsistent, and coordinating the variety of potential different programs to contribute data globally would be a challenge. Each of the different existing data contributions would need to be documented, and the resulting lack of globally uniform, consistent methods may be unsatisfying. Incorporating existing data into a global accuracy assessment merits consideration, but these data are more likely to play a minor role providing limited-purpose supplemental information rather than serving as a panacea for the significant problem of cost of a global accuracy assessment.

3.4. Other Sampling Design Issues

The sampling design for global accuracy assessment should have a built in mechanism permitting easy supplementation of the sample if additional resources become available. The ability to supplement the sample within a large geographic region is particularly desirable in the event that, for example, a country or group of countries provides funding to more precisely estimate accuracy for subregions of the map. As a general rule, the selection and analysis protocols for supplemental sampling are easier if the original sampling design is simple. For example, a stratified random sample is readily augmented by selecting additional sample units from each stratum, and the analysis remains a fairly straightforward application of stratified sampling estimation formulas (Overton and Stehman, 1996). The sampling designs most likely practical for global assessments (*i.e.*, combining two-stage cluster sampling with stratification) need to be evaluated to determine how easy it would be to supplement the sample and to produce accuracy estimates for the supplemented design.

The design and implementation of an accuracy assessment for a global land cover product can be approached from two directions – a uniformly consistent, centralized "top down" approach, or a decentralized, regionally autonomous approach (*e.g.*, a region is a continent). Each approach has advantages and disadvantages.

In the regional approach, each contributing partner determines the sampling and response designs implemented in their region. Regions would not necessarily adopt the same protocols, for example, placing more or less emphasis on existing data or other source materials used for reference data interpretation. Advantages of this approach include the more localized control and sense of ownership of the process, which may translate into better assessments within each region. Administering and implementing the design and analysis may be easier within a regionally-directed framework, and the relationships needed to effectively access and use existing data may be enhanced by this approach. Global mapping efforts implemented with a regional organizational structure should contribute relevant experience to implementing this same approach for accuracy assessment. Cost of integrating the information derived from a regionally designed and implemented assessment may be higher than for a centralized approach because of the considerable effort required to document the greater variety of response and sampling design protocols, and to construct estimates from potentially different regional strategies. Each region would be treated as a stratum in the analysis, so a standard estimation framework exists for combining data from multiple regions into a global assessment.

A more centralized assessment has significant advantages of consistently and uniformly applied protocols, and a globally integrated design may be more likely to yield better precision. Obtaining buy-in from multiple partners to a centralized, top-down plan may be difficult, but if the partners' participation is solicited at the planning stage, this problem may be diminished. A unified approach also has the potential benefit of being able to create an assessment that is of greater overall utility. That is, different but concurrent global mapping projects will all have the need to produce valid accuracy assessments. Multiple groups each independently collecting reference data inefficiently depletes the already limited resources collectively available for accuracy assessment. A well-designed, unified approach to global accuracy assessment may be able to provide high quality reference data, along with a general framework for their use, that are suitable to more than

one global mapping effort. It is less likely that such multi-purpose global reference data would emerge from a more regionally-directed approach.

3.5. Summary

The sampling strategy proposed for global accuracy assessment relies on design-based inference, a widely-accepted statistical foundation for inference in survey sampling. The two key features of the sampling strategy are implementing a probability sampling design and consistent estimation, as guided by rigorous design-based inference. The sampling design will likely incorporate stratification to target objectives of precise class- and region-specific accuracy and clustering to enhance cost-effectiveness of reference data collection. Integrating both strata and clusters into the design is one of the more difficult aspects of global accuracy assessment sampling design. Because the reference data ideally will be made available to the scientific community and will be used for many different accuracy analyses, the challenge is to construct a simple design that is amenable to multiple uses yet still achieves both precision and cost efficiency criteria. Incorporating existing data into a global accuracy assessment merits consideration, but with the recognition that existing data still have costs associated with their use. A general formula has been provided for obtaining consistent estimators of many accuracy measures. This formula (3.2) directly incorporates the inclusion probabilities determined by the sampling design. Reference data error is a potentially significant problem for global accuracy assessments and should be addressed in the analysis stage. No universally recognized standardized approach exists for accommodating reference data error, but some provision should be made in the analysis for these errors, including the option of reporting accuracy results based on several different definitions of agreement.

4. Qualitative Systematic Accuracy Review

Although a statistically rigorous assessment predicated on a probability sampling design is still the "gold standard" for accuracy assessment, other approaches that are less costly can add significantly to the understanding of errors and the potential improvement of the map's accuracy. One of these is systematic quality control, which consists of a quick, qualitative survey that is performed over every part of the map. This systematic assessment of the quality of the maps during and after the classification phase significantly increases the quality of the final products and is recommended as a preliminary step prior to implementing the more formal accuracy assessment.

Systematic quality control arises because recent global land cover products, although of good overall quality, exhibit in some parts major errors that could be avoided by a careful review of the draft products. Such errors reduce the user's overall confidence in the products, even if the quantitative accuracy is high. Errors affecting accuracy of thematic maps can be caused by confusion between the land cover classes (wrong label, missing classes) or can be spatial errors (wrong position of the boundary between classes, disappearance of small patches). The identification of systematic biases affecting some land cover classes or some regions of the world can influence the quantitative validation (sampling strategy) and area estimates.

Systematic quality control is intended to meet two main objectives: the elimination of macroscopic errors and an increase in the overall acceptance of the land cover product by users. Systematic quality control is also a way of assessing if the remotely sensed data have been correctly classified, *i.e.*, if the errors are due to limitations of data quality rather than to poor classification procedures. Systematic quality control should be integrated into the classification procedure, with the results of the analysis employed for removing errors and improving the map.

Accuracy indexes derived from a typical confusion matrix provide information on the quality of the map as a whole but cannot be used to characterize distinct areas of the map. Even when global land cover maps are produced applying the same global algorithm to a homogenous dataset, the quality of the final product is not uniform in all the regions, but instead depends on the quality of observation conditions (cloud coverage, haze, etc.) and ancillary data used to parameterize the classification. In many cases, the land cover map is obtained using a complex classification procedure involving different steps where different algorithms are applied. As a consequence, it is not always possible to derive a per-pixel confidence value as delineated in Section 2.6 and it is necessary to evaluate the accuracy of the results using reference data.

Systematic or regionally stratified sampling schemes (Section 3) allow the computation of accuracy indexes at continental scale, but for some applications users may need spatially explicit estimates of the uncertainty at a finer resolution. The systematic quality control recommended here is a way of estimating the spatial distribution of the errors of a land cover classification.

4.1. Systematic Survey

Qualitative validation is based on a systematic descriptive protocol, in which each cell of the map is visually examined and its accuracy documented in terms of type of error, landscape pattern, reference material used, etc. The grid size cannot be universal but must be adapted to the characteristics of the landscape, the map, and the reference material. For example, in the central part of the Amazon Basin or in the heart of the Sahara, the grid cells can be much larger than in the complex landscapes of Western Europe. A cell size of 200 to 400 km on a side is proposed as a target for providing a good idea of the overall quality of a global product, keeping in mind that the goal of this exercise is a quick survey.

For the systematic survey, different reference materials can be used, including single-date coarse resolution images, detailed thematic maps, and quick-look imagery derived from fine-resolution sensors. Preprocessing and classification procedures applied to multidate imagery often lead to the loss of many spatial details that are clearly visible on coarse resolution single-date images. This

loss of detail is particularly obvious when long time series of derived parameters such as vegetation or moisture indexes are used as input for the classification. The acquisition dates of the single images used for the comparison with land cover maps are crucial for obtaining a reliable assessment. A careful examination of the phenological cycles should be conducted in order to choose the most characteristic period(s) of the year.

Many parts of the globe are covered by detailed thematic land cover, land-use, or vegetation maps that can be used in this quality assessment. These are increasingly derived from earth observation techniques, but benefit from intensive field campaigns and local expertise. One characteristic of these maps is heterogeneity of the legend. Indeed, many maps have definitions of basic classification attributes (dominant layer, height and coverage of the different layers, water regime, etc.) that are too detailed for a clear comparison with global maps, which are typically more generalized. Before the two maps can be compared, the legend must be usually expressed in a common language with similar parameters. Products like Africover (FAO, 2004), with a legend fully documented by hierarchical classifiers, present clear advantages for the comparison of fine- and coarse-resolution maps. It is important to note that the spatial aggregation processes for global land cover products lead to mosaic classes that often do not exist in local maps.

Note that while establishing a correspondence between legends is often a useful step before comparing maps, it is not necessarily essential. For example, Finn *et al.* (1993) demonstrate the use of a mutual information index in comparing maps using different classification schemes.

In the past few years, free access to large fine-resolution datasets has greatly increased. In particular, the years 1999 to 2002 are well covered by freely-available Landsat and SPOT data, thanks to the action of space agencies and international efforts such as GOFC-GOLD. Also, the quick-look images freely available through web portals are often sufficient to characterize the overall quality of a cell, especially for spatially homogeneous areas. Moreover, the management of a large number of small volume quick-look images does not require specific data handling systems. In complex cases, where the spatial fragmentation precludes the use of quick-looks, full resolution images should be used for the quality control. The high resolution or quick-look images, like the coarse resolution ones, suffer from the limitation of being single-date observations, while in some cases multitemporal data series are necessary to discriminate among land cover classes.

4.2. Quality Assessment

In a systematic quality control exercise, each cell examined during the quality control procedure is characterized in detail by a few parameters: the composition and the spatial pattern of the cell, its comparison with other existing global land cover products, the overall quality of the cell, and the nature of any problems.

The cell composition is a key factor affecting the precision of a map because some land cover classes (*e.g.*, evergreen forests, deserts, water bodies) are easier to discriminate than others (*e.g.*, deciduous forests or woodlands, grasslands, extensive agriculture). Information on the composition of the cell contributes to a better understanding of the errors and can help to stratify the population, as in design-based inference. On the other hand, some users focus on specific land cover classes and will be interested in a spatial representation of the errors for cells dominated by their class of interest.

It is widely recognized that the spatial pattern of the landscape influences the appearance or disappearance of land cover classes at varying resolution as well as the area estimates derived from coarse resolution maps (Moody and Woodcock, 1994; Mayaux and Lambin, 1995). Landscape heterogeneity can be expressed by means of qualitative definitions (*e.g.*, highly fragmented, moderately fragmented, little fragmented, not fragmented) or by quantitative metrics (*e.g.*, diversity, perimeter-area ratio, mean patch size). A catalog of qualitative fragmentation categories should be completed before starting the evaluation process in order to insure consistent categorization throughout the map. A basic quantitative estimator of the landscape complexity, like the Shannon entropy index, can and should be computed for every cell. Specific quantitative

metrics of spatial pattern can be also applied. They should be selected on a case-by-case basis, since many indexes are class-specific and can be useful only if proper classes are identified. Computing heterogeneity indexes, as well as reporting the composition of each cell, can be systematically performed in a GIS.

Systematic comparison with existing global land cover products based on remotely sensed data can be also performed for each cell. At least three products (GLC2000, MODIS Land Cover, IGBP-DISCover), derived from data acquired by different sensors, are presently available for a systematic comparison. Because the products adopt different legends, the comparison should be conducted for corresponding groups of classes. A simple agreement measure, like the percentage of pixels with the same label, can be easily computed. Also available is the average mutual information index proposed by Finn (1993), which does not require legends to strictly correspond.

The overall quality of each cell can, as a first approximation, be categorized in qualitative classes using a linguistic scale. As an example, GLC2000 used five classes: excellent, very good, good, moderate, unacceptable. As with qualitative labeling of heterogeneity, a catalog of representative cases should be provided in order to ensure consistency. The labeling of overall quality, once performed for all the cells, allows for a synthetic spatial representation of the quality of the product.

4.3. Nature of the Problems

Ascertaining the nature of the errors occurring in the cell is of primary importance. Statistical accuracy assessment merges in the category “error” many different cases that quality control can easily document. Such information can be profitably used for improving the map during the updating phase. The main cases that can be found in global products are the following:

- The delineation of a land cover feature is accurate, but the label is wrong. In this case, the type of confusion must be specified in order to derive a thematic “distance” between the right and the wrong labels. It is, for example, generally more problematic to classify tropical forests as grasslands than to classify woodlands as savannas.
- The proportions of labels present in the cell are generally correct, but the delineation of the various features is wrong. If this case is the most frequent, it means that the spatial resolution (and eventually the preprocessing steps) precludes any accurate delineation of land cover features. The first global land cover products derived from AVHRR suffered from limitations, such as geolocation. The extreme case of this category occurs when no clear structures appear on the map. The land cover map then corresponds more to a climatic stratification.
- One important land cover feature is missing in the map or a feature is mapped while it is not present in the field. This is a particular case combining a wrong label and an inaccurate delineation of the land cover features. For example, it happens when specific features are derived from erroneous ancillary data, like planned infrastructures never actually built (dams).

Once all the cells have been visited, and the various fields stored in a database, it is possible to investigate the influence of the parameters (heterogeneity, dominant class) on the quality of the land cover map. Some of the interactions that can be investigated are:

- Map quality vs. land cover classes: Is the quality of the map uniform among the different land cover classes?
- Map quality vs. landscape diversity and fragmentation: Is the quality of the map the same in simple and in complex landscapes?
- Map quality vs. agreement with other global land cover maps: Are the errors mainly located in the areas of poor agreement with other maps?

- Land cover classes vs. type of error: Do land cover classes suffer always from the same type of error?

4.4. Comparison with Other Datasets

It is possible to compare the land cover map with existing maps, such as regional and national maps. The comparison must take into account all the issues of compatibility between the datasets, including varying legends (which might require the collapsing of some classes), time of production of the reference maps (land cover changes in the time interval), and scale of the maps (geographic aggregation for deriving area estimates). National census data might be also used as reference data for area estimates, particularly for agricultural areas, that often do not appear as a separate class on reference maps. Attention must be paid to the quality of the census data (being based on self-declaration, they might be biased) and at the geographic level at which they are aggregated.

4.5. Test Sites

In acquiring validation data, the number of sample "cells" (or plots or areas) and intensity of details and accuracy at each location will depend on the resources available. Increasing the number of samples and the accuracy requirements for each sample increases time and labor costs. While a rigorous, globally representative, statistical sample is optimal for quantitative assessment, the cost of such an analysis can be high. However, some information can be gained by a less comprehensive sample.

In order to recognize a continuum ranging from a few samples at low cost to a thorough sample at a higher cost, the Land Product Validation Working Group has established the following validation hierarchy related to the nature and intensity of sampling:

- **Stage 1 Validation.** Product accuracy has been estimated using a small number of independent field measurements obtained from selected locations and time periods.
- **Stage 2 Validation.** Product accuracy has been assessed over a widely distributed set of locations and time periods using a larger number of independent field measurements.
- **Stage 3 Validation.** Product accuracy has been assessed and the uncertainties in the product well established using independent measurements in a systematic and statistically robust way properly representing global conditions. This hierarchy has been established in light of the common practices of global land product producers. Most producers will have the ability to conduct Stage 1 validation as part of the project that is funding the product creation. With some additional effort and possibly some international collaboration, Stage 2 validation can be accomplished with only marginally more infrastructure and resources. While Stage 3 validation requires a significant amount of resources (and consequently, specific funding), it should be noted that important information can be learned in Stage 1 and Stage 2 validation and the results for such can help in setting up a Stage 3 validation study.

4.5.1. CEOS Test-Sites

In an attempt help CEOS members more efficiently reach a Stage 2 validation for their global land product, the Land Product Validation working group is presently establishing a set of CEOS Cal/Val test sites. Building on NASA Earth Observing System's Land Validation Core Sites, the CEOS test-sites are meant to serve as a focal points for validation of multiple global land products (Morisette *et al.*,2002). This focus allows for collaboration within and among science teams and reduces the duplication of effort that would result from validation efforts at disparate sites. Although LPV is leading the development of these sites, they are intended for use by other subgroups within the Working Group on Calibration and Validation. The concept is to build a network of sites where imagery and derived products provided by CEOS members are freely available via the Internet. For land cover, the plan is for producers participating in the CEOS test-site activity to subset their land cover products over the test sites. These data are complemented by

detailed land cover maps derived from high resolution satellite data classified by regional experts. Currently the CEOS test-sites concept is being developed at the stage of a pilot study. There are currently only five sites where global land products and high resolution data are available. Planned activities include expanding the number of sites and working with CEOS members and regional experts to create a high resolution land cover map from the high resolution imagery available at each site.

5. Validation of Global Land Cover Change

5.1 Change versus Single Time-Frame Characterizations

The process of validating a land cover change product has special considerations which make it different from that of an individual land cover characterization. A land cover change accuracy assessment is concerned with the changes between two time periods, as opposed to an instantaneous or time-integrated land cover map (Macleod and Congalton 1998). At the global scale, the complexities arising from this simple change in reference frame can be daunting. First, there is no possibility of deriving a static global set of validation sites, such as might be used in validating a single time-frame land cover map. Land cover change is spatially distributed in a heterogeneous way and dynamic over time. Change events also represent relatively rare cases in time-series land cover mapping efforts, especially so at the global scale. Thus, any simple or stratified random sample which was created to efficiently assess single time-frame global land cover would be inadequate for assessing change classes. If a global validation set for assessing land cover map accuracy were created, it may only be of use to the portion of the change matrix which represented areas not undergoing change and only for the time periods concurrent with the change detection study.

Second, validation information must be gathered at each validation site for both time 1 and time 2 states. At the global scale, the possibility of acquiring such data is compromised by uncertain availability and high cost, certainly double that of a single time-frame classification assessment per site. This added temporal dimensionality also complicates sampling considerations. A change detection validation is not concerned only with the individual cover classes, but with all of the possible from-to land cover change class combinations as well.

Third, the success of global change detection studies is a function of independently derived time 1 and time 2 map characterizations. If the initial products are of inferior quality, then the validation exercise could end up being an investigation of errors found in the input land cover characterizations, not a measure of actual land cover change. As moderate- and coarse-resolution global data sets consist predominately of difficult-to-map mixed pixels, and change typically occurs at subpixel scales, there is reason to believe that the ability to measure change may be limited. The likelihood of successfully using a post-classification approach to change detection at the global scale is suspect. Mapping change with fuzzy measures derived from classification procedures or sub-pixel characterizations of land cover is undoubtedly preferred (see Section 2.5). Regardless of these obstacles, there is a pressing need to quantify global land cover change in a timely manner and new global data sets offer this possibility (Zhan *et al.* 2000; Friedl *et al.* 2002; Bartholomé *et al.* 2002; Hansen *et al.* 2003; Tansey *et al.* 2004). This section outlines the issues germane to change using global land cover data sets derived from remotely sensed imagery.

5.2 Defining Land Cover Change Types

Land use and land cover change includes both the conversion from one land cover category to another (Riebsame *et al.* 1994) and the modification, or subtle within-class change, that affects the character of the land cover without changing its overall classification (Coppin *et al.* 2004). The ability to detect land cover conversions is a function of the mapability of the classes themselves, the spatial extent of change, and the temporal context in which the change occurs (Singh, 1989). Addressing what cover change dynamics are expected to be detected is the first order of business.

From the temporal perspective, land cover change can be ephemeral, interannual or semipermanent/permanent. Ephemeral changes are short-term changes in cover, such as floods or seasonal burning in a savanna setting, which do not permanently alter the dominant vegetation cover distribution of the landscape. Interannual changes are variations in land cover largely due to long-term climatic variability, such as change in the annual extent of grasslands in the Sahel or reduction of woodland canopy cover for an area experiencing long-term drought. Semipermanent/permanent changes are wholesale land cover conversions and include new

construction of impervious surface, deforestation events, or the expansion of agricultural lands. Land cover modifications, as compared to land cover conversions, are a form of semipermanent/permanent change within a given land cover category. This is a more subtle form of change and includes examples such as rangeland degradation due to overgrazing and forest thinning due to selective logging. Using global data sets, all of these types of land cover changes can be detected. However, assessing the accuracy of each type requires a separate validation exercise based on the variation of the temporal dynamics of the change.

Regardless of the change scenario being studied, there normally should be a set of exhaustive and mutually exclusive definitions that describe the various cover states. This premise applies to change detection analyses as well as single time-frame land cover classifications. Physiognomic-structural based definitions are favored for analyzing change at the global scale for several reasons. First, land cover, as defined as the observed biophysical state of the earth's surface, lends itself most unambiguously to a physiognomic definition set (DiGregorio and Jansen, 2001). Second, the signal being mapped with global time series satellite data is highly correlated with vegetation structure and phenology in terms of life form and cover. Third, physiognomic-structural definition sets based on measurable traits such as cover and height allow for validation exercises that can measure these same traits.

Upon constructing a definition set for time 1 and time 2 states, possibly drawn directly from a land cover classification legend, expected algorithm limitations must be examined. What is the feasibility of mapping change given the temporal window (time 1 to time 2) and the spatial resolution of the data sets used in the analysis? For example, annual updates of global forest cover will not account for selective logging due to inadequate spatial detail in the satellite data. Forest regrowth also cannot feasibly be measured over annual increments due to the limited change seen within most regrowing canopies over a single year. Thus, while a stated goal of a change detection study may be to map forest change, distinct and important subcategories of this cover conversion type may be wholly absent from the analysis. As such, statements of change regarding forest cover must account for the percentage of forest change that is represented by these subcategories, either through a literature review or a more intensive validation exercise. By accounting for these subcategories in this way, needs of the user community in applying the data are better addressed.

5.3 Change Accuracy Assessment Using Categorical Data

For discrete characterizations of land cover, the most widely used framework in assessing map accuracy has been the confusion, or error matrix (Congalton, 1991; see Section 2.3). The dimensions of a single time-frame land cover classification error matrix of N classes are simply $N \times N$. However, when comparing consecutive land cover depictions, a new matrix representing all of the from-to thematic conversions must be created. Considering all possible conversions, the dimensions of the change detection error matrix become $N^2 \times N^2$ (Khorram, 1999). Figure 5.1 shows the situation for a three-class case. The greater dimensionality of the error matrix leads to increasing complexity and associated costs concerning sampling procedures in order to fill the matrix.

One way to simplify the problem is to collapse classes that are either of no interest to users of the data, are unlikely in nature to form from-to change categories, or are not likely to be detected. Figure 5.2*a* and 5.2*b* show the University of Maryland global land cover classification scheme and how it can be reduced to a subset of categories better related to global-scale land cover change dynamics. While this subset may still represent a complicated sampling frame, it has greatly simplified the original reference legend. This subset represents the cover change classes used by the University of Maryland 250-meter change product derived from MODIS data (Zhan *et al.* 2000).

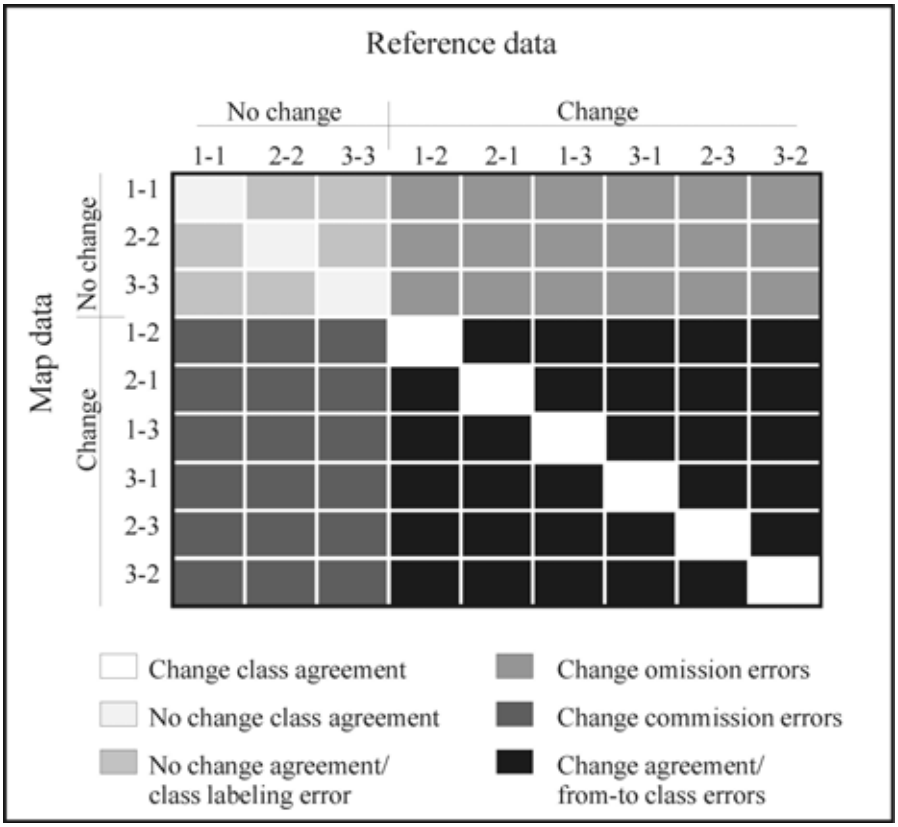


Figure 5.1. Change detection confusion/error matrix for time 1 and time 2 maps each consisting of 3 land cover classes

Figure 5.2*c* shows a further simplification by taking selected themes from the subset categories to create change/no change matrices for an individual land cover conversion process, in this example deforestation/afforestation. This particular case is further simplified in Figure 5.2*d* by aggregating the change classes in order to create a binary case for two categories, deforestation and not deforestation. This simplification could be performed for any change subset over any desired time interval, such as agricultural expansion over a 5-year period. In so doing, individual assessments of specific land cover change scenarios can be examined in a binary mode, simplifying the sampling procedure. For evaluating a global land cover change map, we advocate using a simplified binary model because of the increases in sampling efficiency and the reductions in overall costs. Selecting which themes to individually validate would be driven by user priorities and available funding. The use of confusion/error matrices in deriving overall and per category accuracy estimates is described in Section 2.

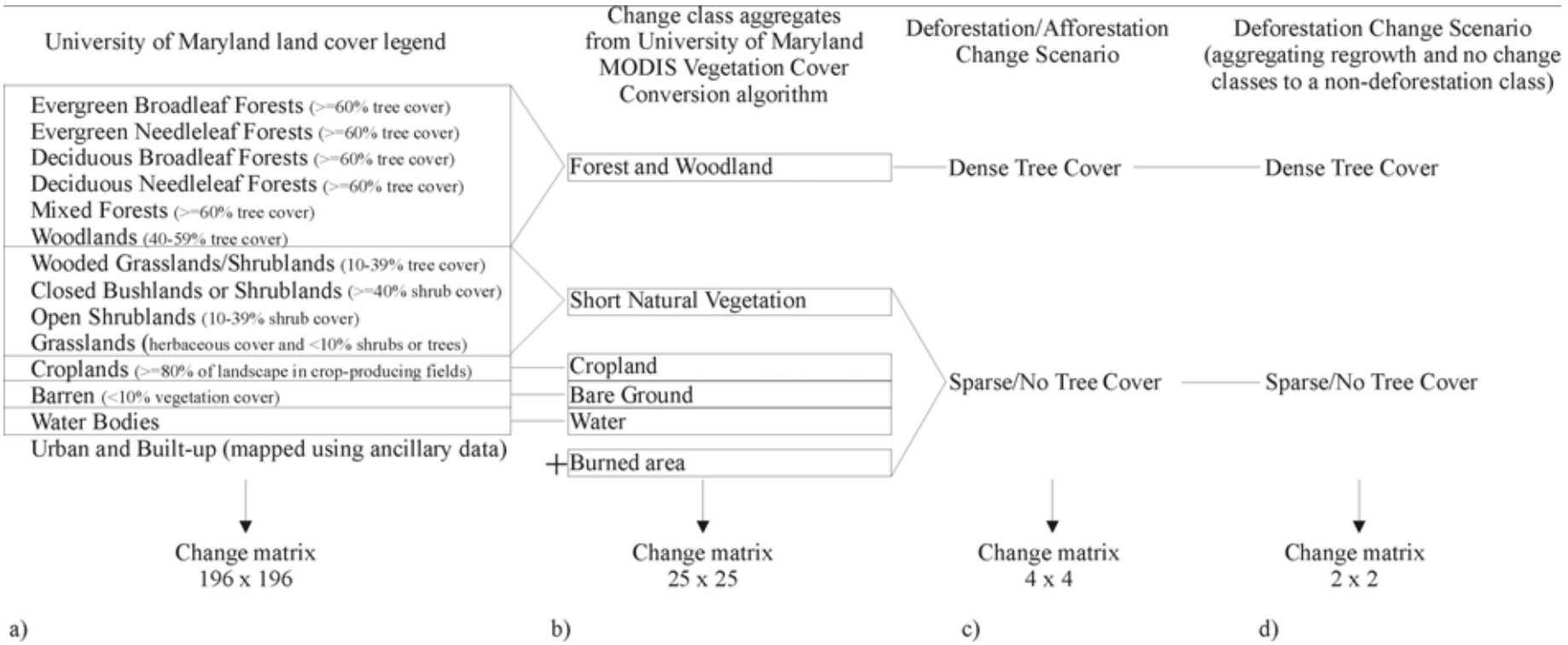


Figure 5.2. a) University of Maryland land cover legend, b) Aggregated classes for detection of land cover change events at the global scale (Zhan *et al.* 2000), c) Simplified change scenario for validating deforestation/afforestation, with from-to classes consisting of dense to dense, dense to sparse, sparse to sparse and sparse to dense tree cover d) Two category change scenario further simplified by aggregating dense to dense, sparse to sparse and sparse to dense into a single class

5.4 Change Accuracy Assessment Using Continuous Representations of Land Cover

Measuring the accuracy of change detection maps derived from successive continuous representations of land cover is in some ways simpler than that derived from consecutive discrete classifications. Continuous map layers have numerical output attributes which allow for the use of more traditional, and often more intuitive, measures of accuracy assessment, such as root mean square or standard error values. A particular strength of mapping land cover in a continuous fashion is the possibility of detecting subpixel land cover change. While consecutive classifications are used only for the identification of from-to categorical change, continuous cover maps can also detect within-class change, revealing within-cover-type modifications to the land surface. Continuous cover maps, in fact, are free from categorical definition sets, being measures of the presence or absence of basic vegetation types such as tree cover or crop cover. Change can be measured using a simple difference imaging approach and such an approach has been used at the global scale for estimating deforestation (Hansen and DeFries, 2004).

Because explicit numerical cover estimates imply a higher degree of precision than categorical cover labels, they also demand more precision from reference data. In this regard, continuous cover validation efforts are more demanding than categorical change assessments. Datasets such as Landsat Enhanced Thematic Mapper Plus images, which may be used to identify broad land cover categories such as forest and bare ground, cannot as easily be used to estimate percent cover. Even ground-based methods for validating land cover classes, such as drive-by field sampling, are not precise enough to be used in assessing the accuracy of continuous cover maps. While research has been performed in developing validation methods for single-time frame continuous cover maps (DeFries *et al.*, 2000), there has not yet been an effort made to validate continuous cover change estimates. The obvious difficulty is deriving reliable time 1 cover estimates given the probable lack of available data, such as in *situ* field measurements or very high resolution satellite imagery. Given the likelihood that subpixel numerical attributes, whether fuzzy class probabilities or percent cover estimates, are the future of global land cover change mapping, additional research in this topic area is required.

5.5 Sampling the Map for Assessing Change Detection Accuracy

Sampling for measuring the accuracy of a change detection product is driven by the fact that the change category is exceedingly rare. While change could be a common feature in the landscape for studies of smaller areas, we will assume that it is always an isolated occurrence at the global scale. Figure 5.3 shows a potential scenario in global change detection. This is a reasonable scenario as deforestation (change from forest to nonforest) is often an abrupt and spatially dramatic event while reforestation (nonforest to forest) is longer term and may be ignored in a short-term inventory. It is clear that in order to state each category's accuracy within a suitable confidence level, some sort of stratified sampling protocol or other accommodation must be administered for each change class.

Stratification of the map for performing an accuracy assessment may be achieved simply by identifying the change and no-change pixels. However, this approach does not address the case, particularly if it is rare, of errors of omission which would likely not be quantified in a "no-change" class map stratum. At the global scale, the correctly classified no-change category should dominate the no-change stratum as shown in the Figure 3 example, to the extent that a sample taken within it will not reveal the true extent of false negatives. In order to improve the sampling of strata to account for errors of omission, the strata may be modified. Areas of change from the map itself can be expanded or areas known to be experiencing high rates of change via domain knowledge can be added to the sampling scheme (Khorram, 1999). In this way, areas of likely change not flagged as change are added for use in assessing the occurrence of errors of omission.

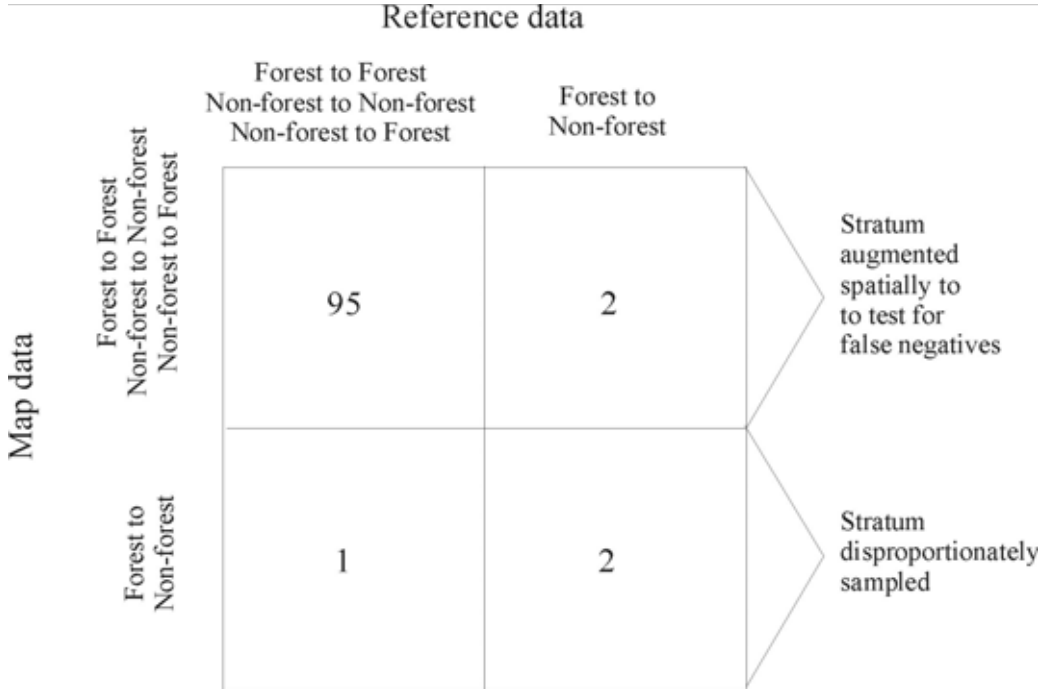


Figure 5.3. Example change matrix for measuring deforestation and nondeforestation classes at the global scale. Values of each cell are normalized by the sum of all cells.

For validating change detection maps, it has been advocated that researchers employ a stratified disproportionate sampling scheme. After stratifying the map into change and no-change classes, the change stratum is then disproportionately sampled. Disproportionate sampling is a design which increases the fraction of samples taken within the stratum where the rarely occurring cases are concentrated (Biging *et al.* 1998). The primary benefits of the approach are the increased precision in estimating accuracy of the change category (user's accuracy) and the more efficient use of resources in collecting reference data within areas more likely to have experienced land cover change. With respect to estimating overall accuracy, gains using disproportionate sampling over proportionate sampling occur when the overall thematic accuracy of the map is high (greater than 80 percent), when the change stratum represents 15–25 percent of the study area, and when user accuracies in the change stratum are low (Biging *et al.* 1998). Gains in relative precision for overall accuracy are minimal when the change stratum is very rare.

For global applications where the change stratum should represent considerably less than 15-25 percent of the land surface on an annual to 5-year basis, disproportionate sampling may not yield a marked increase in sampling efficiency for estimating overall accuracy, although it will still benefit precision for estimating accuracy of the change class. If change is very accurately mapped, stratification based on map change will function equally well as a stratification for reference or true change, and accordingly improve precision of the estimated producer's accuracy for the change class. If change is not mapped accurately, the sample size for true change may be very small even with disproportionate stratified sampling based on map change. Ensuring an adequately large sample of true change pixels is difficult if the change class is exceedingly rare, or if the identification of such change with reference data is prohibitively expensive, and no solution to the sampling problem may exist (Kalton and Anderson, 1986).

For global change analyses, polygon sampling should yield a higher efficiency in sampling unique change events. Most global validation work has been, and will be, performed on individual pixels. This is primarily due to the fact that moderate- and coarse-resolution pixels are typically mixtures of land cover types. As a consequence, there is no expectation of deriving homogeneous, thematically coherent polygons. While most change events derived using such data should occur at the individual pixel scale, many single change events cover multiple pixels. Examples include

forest fires and conversion of natural cover to mechanized cropland. These change events co-located in time can be polygonized and used to improve sampling efficiency. For sampling procedures, it is important to account for the inclusion probabilities whether the approach is based on a per-pixel or polygon sampling basis (Stehman and Czaplewski, 1998).

The approach to global land cover accuracy assessment described in Section 3.1, where both clusters and strata are used to spatially constrain validation study areas, is equally applicable to global change validation. Using Landsat footprints as a sampling frame in concert with different forest cover strata has been advocated by Czaplewski (2002) and employed by Achard *et al.* (2002) to estimate continental-scale forest change. Czaplewski (2003) also demonstrated the utility of Landsat as a sampling basis for global forest change estimations. The Achard *et al.* study used high-resolution imagery as clusters in conjunction with a nonchanging forest stratum and a more intensively sampled "hotspot" of deforestation stratum to estimate tropical deforestation from 1990 to 1997. These approaches could easily be translated into a global land cover change validation methodology. However, little research concerning this topic has been performed. At the global scale, the derivation of sampling schemes optimized for validating land cover change has not been attempted or thoroughly examined.

5.6 Algorithm Level Confidence Measures

At the algorithm level, the first assessment of the quality of a land cover change product involves the biophysical modeling of change trajectories in spectral space. Modeling the spectral movement through time of various change scenarios can yield a qualitative statement about candidate change sites. These trajectories could also be compared to time-sequential spectral measures which have been shown to be related to land cover change dynamics. Trends in vegetation greenness (NDVI) and land surface temperature (Lambin and Ehrlich, 1996) could be correlated with the change product in order to provide a biophysical basis for assigning confidence to change sites.

A more typical approach employs training data to assign confidence levels to change pixels. The degree to which the calibration information is correctly modeled is an indication of the quality of the final product. For global land cover change mapping exercises, specifically-labeled change training data are not used, as such data are not readily available. Instead, change is found via the comparison of successive land cover characterizations. Time 1 and time 2 confidence measures can be used to derive a combined metric for assigning the likelihood of change per pixel. In estimating global forest change from 1984 to 1997, Hansen and DeFries (2004) used training data from areas not experiencing change to delineate various thresholds for identifying likely change pixels. From this, a relative confidence of change per pixel was created. Assigning confidence in this way is simple to perform, and should be used for global change studies in the absence of independently derived validation data.

5.7 Independently-Derived Reference Data

There are many potential sources of independently-derived land cover change information. The typical limitation of such data sources is the lack of a systematically-derived sample. While probability sampling is the goal of any validation exercise, at the global scale this is a difficult and costly aspiration. This section reviews possible data sources for assessing the quality and/or the accuracy of a derived global land cover change map, regardless if probability sampling is used or not.

While *in situ* ground measurements may be a preferred source of validation data, their use at the global scale in assessing the accuracy of land cover change maps is limited due to high costs. If it were possible to incorporate field visits, they would be most useful for describing time 2 conditions. On the ground, time 1 land cover state would have to be divined using surveys of local landholders, an impractical and costly scenario at the global scale. Instead, any field data obtained would have to be used in conjunction with archival imagery to make a statement about time 1 land

cover conditions. For some studies, ground information may be a necessary requirement for successfully assessing accuracy.

Intermediate-scale satellite data and derived map products are useful data sources for global change validation efforts. In fact, high-resolution satellite data may be more useful for a change product validation than for a single time-frame characterization, due to the often dramatic spectral variations seen in land cover change events. For example, identifying the differences between similar cover types, such as broadleaf evergreen and broadleaf deciduous forests, may well be impossible given a single Landsat image. For a global change product mapping the cover transformations outlined in Figure 5.*2b*, individual Landsat images should be of greater utility. Of course, the use of even finer spatial detail imagery, such as IKONOS data, would provide even higher confidence in validating change. However, data costs and availability of such imagery limit their use in validating land cover change at the global scale. Of crucial importance is assembling reference data sets that are directly comparable to the land cover change maps. High-resolution data sets must be converted to thematic classes congruent with the land cover change definition sets. Also, validation data need to be acquired which conform temporally to the global map products, which typically represent a time-integrated annual land cover signal.

Existing land cover change maps at the same scale or finer can also be used to assess product quality. The primary limitation of such an approach is the rarity of such data sets in space and time. Examples include the Pathfinder humid tropical deforestation project data sets (Townshend *et al.* 1995), which mapped decadal change for areas of the Amazon and Congo basins, and the PRODES data sets from the National Institute for Space Research (INPE) in Brazil, which depict annual deforestation in the Legal Amazon (INPE 2003). Assemblages of such data have been used to validate and calibrate global forest change estimates from continuous tree cover change maps (Hansen and DeFries, 2004). When incorporating these data into a validation exercise for a global change product, attention must be paid to the land cover definitions employed. Equally problematic is the variable timing of the mapped change events. For example, if the global data set represents a March to March annual map of land cover while the high-resolution reference change maps have variable time intervals, then direct comparability is limited. Such reference data must be filtered to find the source information that best conforms to the timing of the global-scale change product.

An important source of validation information is the intercomparison of global change products mapped at similar spatial resolutions. An example of this is the possible comparison of global estimates of annual burned area. Currently, the Joint Research Center of the European Commission is mapping global burned area using SPOT VEGETATION data (Tansey *et al.*, 2004), while NASA's MODIS Land Group is implementing a new approach to mapping this same theme (Roy *et al.*, 2002). Testing the thematic concurrency between products derived from different sensors and algorithms would be a valuable validation exercise.

Inventory data that document change using non-site-specific estimates can also be used as reference information for validating land cover change maps. The area of study is typically an administrative unit such as a state or country within which changes in cover are tabulated. The major limitations to using inventory estimates are (1) data inconsistency due to variations in data collection methodologies across study areas; (2) variable timing of data collection; (3) incompatible definition sets; and (4) data quality issues. Data on national scale forest area change from the United Nations Food and Agricultural Organization's Forestry Resource Assessment (FAO, 2000) have been compared with global change estimates derived from synoptic satellite data (Hansen and DeFries, 2004). Other change information sources such as agricultural databases could likewise be used in inventory comparisons.

6. Recommendations and Conclusions

Accuracy assessment is an expensive, yet essential, component of the land cover mapping process. Maps without associated accuracy data remain untested hypotheses. All projects funded by CEOS member organizations should require accuracy assessment of all maps produced for use by the wider scientific community. Statistically valid estimates of map accuracy and their publication are essential to validation of land cover products and their ultimate acceptance and use.

A set of core analysis methods exists for accuracy assessment that should be routinely adopted as a baseline for reporting map accuracy. These include employing probability sampling and consistent estimators within the design-based inference framework to generate estimates of the overall accuracy of the map as well as per-class accuracies and the variances of these estimates. Confusion matrices, user's and producer's accuracies should be published with the accuracy assessment, and the data used to derive these estimates should be archived and made accessible to the scientific community.

There is considerable room for building upon these core methods to pursue additional dimensions of map accuracy that can improve the validation of land cover products. These are not limited to, but include: (1) validation both during and after map production; (2) use of confidence-based quality assessment methods for assessing spatial uncertainty; (3) addition of fuzzy accuracy methods; (4) appropriate use of systematic qualitative and descriptive methods; and (5) appropriate extensions of single-date approaches to land cover change validation.

Global land cover maps at coarse resolution pose unique challenges for accuracy assessment, including the high frequency of mixed pixels, difficulty in precise geolocation of map products and reference materials, and in the difficulties associated with acquiring and interpreting fine-resolution reference imagery.

6.1. Areas of Future Research

While there is a well-established core set of methods for accuracy assessment of thematic maps, there remains considerable need for future research and development. Areas of particular importance in this domain include:

- *Standardization of land cover maps with respect to legends and mapping units.* To date, most land cover maps have been made independent of existing maps and other mapping efforts. As a result it has proven difficult to compare and combine alternative land cover maps. Efforts to standardize land cover legends and the nature of the mapping units would greatly enhance the synergy between mapping efforts and prove beneficial to the science community.
- *Development of methods for validation of more continuous measures of land cover.* A number of land cover maps now use continuous measures of surface properties, such as percent tree cover, rather than categories of land cover. The existing core methods for assessing the accuracy of thematic maps are not necessarily well suited to these new products.
- *The effect of spatial aggregation on accuracy estimates*. Many users of land cover maps require spatially aggregated products and it is difficult to know the accuracy of these products even if accuracy assessment has been done on the maps that were aggregated. Methods for estimating the accuracy of spatially aggregated products from accuracy assessments at finer resolutions are needed.
- *Reuse of existing validation samples.* Accuracy assessment is expensive primarily because of the costs associated with the validation samples. Thus, there is a strong motivation to use existing data collected for other purposes, and these data are typically difficult to incorporate in a design-based inference framework. Although theory exists to show how new and existing accuracy observations can be merged (see Section 3.3), more work is

necessary to demonstrate the concept. The reuse of existing validation samples might significantly reduce the cost of future accuracy assessment efforts.

- *Validation of change maps.* The validation of change maps poses new challenges and development of new methods is required. In the domain of validation of change detection there is considerable need for development of methods for separating land cover conversion from interannual variability in ecosystem response to climate variability. Integrating accuracy assessment of change with accuracy assessment of single-date land cover maps is a critical need for global monitoring of status and trends in land cover.
- *How to best use spatially-distributed confidence-based metrics in conjunction with conventional accuracy metrics.* Many algorithms for land cover mapping now provide spatial estimates of uncertainty in derived land cover information. These data have been demonstrated to help explain spatial patterns in land cover accuracy. However, more research is needed to formally link spatial uncertainty data, such as confidence values provided by a classification algorithm, with design-based sampling methods to better characterize map accuracy.
- *Misregistration, mixed pixels, and PSF effects.* Particularly for coarse resolution imagery, the problems of misregistration, mixed pixels, and the underlying effective point spread function (PSF) of the sensor confound the accuracy assessment process. More research is needed to better characterize and understand these effects as they relate to accuracy assessment at coarse resolutions.
- *Integrating the effect of error in the reference data*. Conventional methods assume that the reference data ("ground truth") for the sample sites is accurate. It would be desirable to be able to estimate the effect of a known rate of error in the sample sites on the overall accuracy of a map. Although a number of approaches to this problem have been explored in the literature (see Section 3.4), practical techniques to accommodate reference data error remain to be devised.
- *Error magnitude effects*. Conventional methods treat all errors as equal in magnitude, which is clearly not true. Better methods for quantifying the importance of the various types of errors that occur in land cover maps (*i.e.*, fuzzy accuracy assessment) would provide valuable additional information to the science community.
- *Better understanding of users' needs for accuracy data.* An improved understanding of the ways in which the science community uses land cover accuracy data would enhance the ability of future accuracy assessment efforts to provide the most useful information possible.
- *Define priorities for improvements in land cover mapping.* For the purposes of providing guidance to the CEOS space agencies in support of international conventions requiring accurate and detailed land cover data, it would be beneficial to determine where future investment would be most beneficial for improving future land cover maps.

6.2. Toward a Universal Validation Dataset

International processes on the strategic level such as the Global Observation System of Systems (GEO, 2005) and the implementation frameworks for UN conventions (e.g. GCOS, 2004) urge developments towards operational land observations, including a robust and sustained land cover product accuracy assessment. In fact, the most advanced plan for implementing UN conventions (GCOS, 2004) tasks the Working Group on Calibration and Validation of the Committee on Earth Observing Satellites (CEOS-WGCV) and the Global Observation of Forest Cover-Global Observation of Land Dynamics activity (GOFC-GOLD) with outlining reliable and accepted methods for land cover map accuracy assessment, and the development of a sustained in situ reference network with the application of standardized validation protocols. These activities are strongly linked with evolving land cover harmonization initiatives. Harmonization and validation are parallel efforts towards interoperability, product synergy, and improved usability of land cover

products (Herold et al., in press). Overall, this process will improve the value of existing and future land cover datasets for a multitude of applications and contributes to the goal of operational terrestrial observations.

A coordinated international effort and comprehensive consensus building are essential for such a task to be successful. The general approach is to combine experience and resources from different participating agencies involved in global earth observations since previous efforts have suffered from a lack of funding and limited available resources. It seems foolish and wasteful to have each land cover mapping project conduct its own expensive, yet still likely inadequate, accuracy assessment. Instead, the objective is to develop a "universal validation dataset" – a new set of land cover reference sites that provides statistically robust, consistent, harmonized, updated, and accessible reference information that will build upon the validation standards defined in this document.

The universal validation dataset would be based on a core centralized probability sample design. The design would have a moderate level of stratification (e.g., 5 regions by 6 land-cover classes), and it would have built in protocols for how the sample can be supplemented by region or by class. Flexibility to augment the base sample is crucial, e.g. if new datasets evolve or if regional validation activities are to be embedded.

The reference data (response design) development is based on fine-resolution satellite observations. Assuming continuity of satellite observations on this scale, the validation sites will be maintained as "living" land cover reference database that could be used to verify any existing and new land cover map. The interpretations would need to be based on generic descriptions of land cover characteristics in a common language from an internationally agreed classification system such as the UN Land Cover Classification System (LCCS), hence independent of any specific land cover legend. This makes the validation process transparent, consistent, and applicable to any land cover map compatible with LCCS. Understanding semantic differences in existing legends and LCCS translations (as provided by the harmonization activities) are essential for such an implementation and the comparative analyses of accuracy. The reference interpretations have to facilitate the different spatial resolutions of global datasets. From the semantic perspective, LCCS allows the integration for *in situ*/local, regional, and global land cover observations.

The validation analysis activities should focus on several levels of validation using various accuracy reporting measures. The new universal dataset will provide a consistent *primary validation* for all existing global land cover products, supplementing the completed initial validations of the data providers. *Comparative validation* of existing products will be based on comparisons of appropriate accuracy measurements. The comparative assessment is essential for contrasting and comparing different datasets and the development of an interoperability strategy. The basic goal is to identify strengths and weaknesses of individual datasets relative to other land cover products. A comparative validation might also include regional land cover datasets. Given a regular update of the reference database, an operational and continued assessment of the accuracy and validity of datasets can be established even after many years of their production through updated validations. Similarly, for any new global land cover product developed based on LCCS, the universal reference dataset is designed to provide accuracy assessment rigorous comparison between the new product, the ground reference data, and any previous land cover map. As pointed out in section 5, the validation of land cover change has special considerations. Although the proposed stratified sample design is not specifically tailored to focus on a statistically robust validation of change on global scales, there are aspects of such comparisons that have to be considered.

A universal validation set would not necessarily answer all questions adequately, at least not in its initial phase. However, it would provide a strong baseline sample that would meet the need for broad general accuracy statements and serve as the base for an operational land cover validation system. The political framework, the organizations for international cooperation and the methodological resources to support a joint harmonization and validation initiative for land cover

datasets seem to be in place. It is now up to the individual members of the community to provide their share in this initiative. Operational agencies such as the CEOS-WGCV, GOFC-GOLD with its regional networks of local remote sensing/validation experts, and the database management and access systems of the Global Terrestrial Observation System of the United Nations (GTOS) and the UN Global Land Cover Network (GLCN) should play a leading role in development, implementation, maintenance, and dissemination of such a universal database.

7. Literature Cited

Achard, F., H. D. Eva, H.-J. Stibig, P. Mayaux, J. Gallego, T. Richards and J.-P. Malingreau (2002). "Determination of the world's humid tropical forests." *Science* **297**: 999-1002.

Atkinson, P. M., M. E. J. Cutler and H. Lewis (1997). "Mapping sub-pixel proportional land cover with AVHRR imagery." *International Journal of Remote Sensing* **18**(4): 917-935.

Bartalev, S. A., A. S. Belward, D. V. Erchov and A. S. Isaev (2003). "A new SPOT4-VEGETATION derived land cover map of Northern Eurasia." *International Journal of Remote Sensing* **24**(9): 1977-1982.

Bartholomé, E., A. S. Belward, F. Achard, S. Bartalev, C. Carmona-Moreno, H. Eva, S. Fritz, J.-M. Grégoire, P. Mayaux and H.-J. Stibig (2002). Global Land Cover Mapping for the Year 2000—Project Status November 2002. Publications of the European Communities, EUR 20524 EN, Luxembourg, European Commission, 55 pp.

Biging, G. S., D. R. Colby and R. G. Congalton (1998). "Sampling systems for change detection accuracy assessment, remote sensing change detection." In: *Environmental Monitoring Methods and Applications*. R. S. Lunetta and C. D. Elvidge, Eds. Chelsea, Michigan, Ann Arbor Press, pp. 281-308.

Boschetti, L., S. P. Flasse and P. A. Brivio (2004). "Analysis of the conflict between omission and commission in low spatial resolution dichotomic thematic products: The Pareto Boundary." *Remote Sensing of Environment* **91**: 280-292.

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

Campbell, W. G. and D. C. Mortenson (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.

Carmel, Y. (2004). "Aggregation as a means of increasing thematic map accuracy." In: *GeoDynamics: Modelling Spatial Change and Process*. P. M. Atkinson, G. M. Foody, S. E. Darby and F. Wu, Eds. Boca Raton, FL, CRC Press, 440 pp.

CNES (2000). "SPOT-Vegetation Instrument." Centre National pour l'Etudes d'Espace, http://spot4.cnes.fr/spot4_gb/, 06/06/2000.

Cochran, W. G. (1977). *Sampling Techniques, Third Edition*. New York, John Wiley & Sons 428 pp.

Cohen, W. B., T. K. Maiersperger, Z. Q. Yang, S. T. Gower, D. P. Turner, W. D. Ritts, M. Berterretche and S. W. Running (2003). "Comparisons of land cover and LAI estimates derived from ETM plus and MODIS for four sites in North America: a quality assessment of 2000/2001 provisional MODIS products." *Remote Sensing of Environment* **88**(3): 233-255.

Comber, A. J., P. F. Fisher and R. A. Wadsworth (2004). "Identifying land cover change using a semantic statistical approach." In: *GeoDynamics: Modelling Spatial Change and Process*. P. M. Atkinson, G. M. Foody, S. E. Darby and W. F., Eds. Boca Raton, FL, CRC Press, 440 pp.

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. (1998). "Using spatial autocorrelation analysis to explore the errors in maps

generated from remotely sensed data." *Photogrammetric Engineering and Remote Sensing* **54**: 593-600.

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

Congalton, R. G. and K. Green (1999). *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. Boca Raton, Lewis Publishers 160 pp.

Congalton, R. G., R. G. Oderwald and R. A. Mead (1983). "Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques." *Photogrammetric Engineering and Remote Sensing* **49**: 1671-1678.

Coppin, P., I. Jonckheere, K. Nackaerts, B. Muys and E. Lambin (2004). "Digital change detection methods in ecosystem monitoring: A review." *International Journal of Remote Sensing* **25**(9): 1565-1596.

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

Czaplewski, R. L. (2002). On Sampling for Estimating Global Tropical Deforestation. Forest Resources Assessment Working Paper 60, Rome, United Nations Food and Agriculture Organization, 12 pp.

Czaplewski, R. L. (2003). "Can a sample of Landsat sensor scenes reliably estimate the global extent of tropical deforestation?" *International Journal of Remote Sensing* **24**: 1409-1412.

Czaplewski, R. L. and G. P. Catts (1992). "Calibration of remotely sensed proportion or area estimates for misclassification error." *Remote Sensing of Environment* **39**: 29-43.

de Bruin, S. (2000). "Predicting the areal extent of land cover types using classified imagery and geostatistics." *Remote Sensing of Environment* **74**: 387-396.

DeFries, R. S., M. C. Hansen and J. R. G. Townshend (2000). "Global continuous fields of vegetation characteristics: A linear mixture model applied to multi-year 8 km AVHRR data." *International Journal of Remote Sensing* **21**(6-7): 1389-1414.

DeFries, R. S. and S. O. Los (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., J. R. G. Townshend and M. C. Hansen (1999). "Continuous fields of vegetation characteristics at the global scale at 1-km resolution." *Journal of Geophysical Research-Atmospheres* **104**(D14): 16911-16923.

DiGregorio, A. and L. J. M. Jansen (2001). Land Cover Classification System (LCCS): Classification Concepts and User Manual for Software Version 1.0., Rome, United Nations Food and Agricultural Organization, 194 pp.

EEA (1995). CORINE Land Cover, Part I: Methodology, European Environment Agency, 94 pp.

ESA (2004). "MERIS Introduction." European Space Agency, http://envisat.esa.int/instruments/meris/, 07/22/04.

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

FAO (2000). *Global Forest Resources Assessment, FAO Forestry Paper no. 140*. Rome, FAO pp.

FAO (2004). "FAO—The Africover Initiative." Food and Agricultural Organization of the United Nations, http://www.africover.org/africover_initiative.htm

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

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. (2000*a*). "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. (2000*b*). "Estimation of sub-pixel land cover composition in the presence of untrained classes." *Computers and Geosciences* **26**: 469-478.

Foody, G. M. (2002). "Status of land cover classification accuracy assessment." *Remote Sensing of Environment* **80**: 185-201.

Foody, G. M. (2005). "Local characterization of thematic classification accuracy through spatially constrained confusion matrices." *International Journal of Remote Sensing* **26**(6): 1217-1228.

Foody, G. M., N. A. Campbell, N. M. Trodd and T. F. Wood (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. and M. R. M. Embashi (1995). "Mapping despoiled land cover from Landsat Thematic Mapper imagery." *Computers, Environment and Urban Systems* **19**: 249-260.

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

Friedl, M. A., D. K. McIver, J. C. F. Hodges, X. Zhang, D. Muchoney, A. H. Strahler, C. E. Woodcock, S. Gopal, A. Schneider, A. Cooper, A. Baccini, F. Gao and C. Schaaf (2002). "Global land cover from MODIS: Algorithms and early results." *Remote Sensing of Environment* **83**(1-2): 287-302.

Friedl, M. A., A. H. Strahler, X. Zhang and J. Hodges (2003). *The MODIS land cover product: Mapping global land cover properties and dynamics from multitemporal MODIS observations*. International Geoscience and Remote Sensing Symposium 2003, Toulouse, France, IEEE.

Friedman, J., T. Hastie and R. Tibshirani (2000). "Additive logistic regression: A statistical view of boosting." *Annals of Statistics* **28**(2): 337-374.

Gallego, F. J. (2004). "Remote sensing and land cover area estimation." *International Journal of Remote Sensing* **25**: 3019-3047.

Gao, F., C. B. Schaaf, A. H. Strahler, Y. Jin and X. Li (2003). "Detecting vegetation structure using a kernel-based BRDF model." *Remote Sensing of Environment* **86**(2): 198-205.

GCOS (2004). Implementation Plan for the Global Observing System for Climate in Support of the UNFCCC, WMO Technical Document No. 1219, Geneva, WMO, 153 pp.

GEO (2005). "The Global Earth Observation System of Systems (GEOSS)—10-Year Implementation Plan and Reference Document." http://earthobservations.org

Gerten, D., S. Schaphoff, U. Haberlandt, W. Lucht and S. Sitch (2004). "Terrestrial vegetation and water balance—hydrological evaluation of a dynamic global vegetation model." *Journal of Hydrology* **286**(1-4): 249-270.

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

Gorte, B. and A. Stein (1998). "Bayesian classification and class area estimation of satellite images using stratification." *IEEE Transactions on Geoscience and Remote Sensing* **36**(3): 803-812.

Green, E. J. and W. E. Strawderman (1994). "Determining accuracy of thematic maps." *The Statistician* **43**: 77-85.

Hagen, A. (2003). "Fuzzy set approach to assessing similarity of categorical maps." *International Journal of Geographical Information Science* **17**: 235-249.

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

Hansen, M. C. and R. S. DeFries (2004). "Detecting long-term global forest change using continuous fields of tree-cover maps from 8-km advanced very high resolution radiometer (AVHRR) data for the years 1982-99." *Ecosystems* **7**(7): 695-716.

Hansen, M. C., R. S. DeFries, J. R. G. Townshend, M. Carroll, C. Dimiceli and R. A. Sohlberg (2003). "Global percent tree cover at a spatial resolution of 500 meters: First results of the MODIS vegetation continuous fields algorithm." *Earth Interactions* **7**(10): 15 [online journal].

Hansen, M. C., R. S. DeFries, J. R. G. Townshend, L. Marufu and R. Sohlberg (2002). "Development of a MODIS percent tree cover validation data set for Western Province, Zambia." *Remote Sensing of Environment* **83**(1-2): 320-335.

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

Herold, M., C. Woodcock, A. DiGregorio, P. Mayaux, A. Belward, J. Latham and C. C. Schmullius (2005). "A joint initiative for harmonization and validation of land cover datasets." *IEEE Transactions on Geoscience and Remote Sensing*: (In press).

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

INPE (2003). "PRODES: Assessment of Deforestation in Brazilian Amazonia." Instituto Nacional de Pesquisas Espacias, http://www.obt.inpe.br/prodes/index.html

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

Ju, J. C., E. D. Kolaczyk and S. Gopal (2003). "Gaussian mixture discriminant analysis and sub-pixel land cover characterization in remote sensing." *Remote Sensing of Environment* **84**(4): 550-560.

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

Justice, C. O., E. Vermote, J. R. G. Townshend, R. Defries, D. P. Roy, D. K. Hall, V. V. Salomonson, J. L. Privette, G. Riggs, A. Strahler, W. Lucht, R. B. Myneni, Y. Knyazikhin, S. W. Running, R. R. Nemani, Z. Wan, A. R. Huete, W. van Leeuwen, R. E. Wolfe, L. Giglio, J.-P. Muller, P. Lewis and M. J. Barnsley (1998). "The Moderate Resolution Spectroradiometer (MODIS): Land remote sensing for global change research." *IEEE Transactions on Geoscience and Remote Sensing* **36**: 1228-1249.

Kalton, G. and D. W. Anderson (1986). "Sampling rare populations." *Journal of Royal Statistical Society* **149**: 65-82.

Khorram, S. (1999). *Accuracy Assessment of Remote Sensing-Derived Change Detection*. Bethesda, Md., American Society for Photogrammetry and Remote Sensing Monograph Series, 65 pp.

Kish, L. (1965). *Survey Sampling*. New York, John Wiley & Sons, 643 pp.

Kott, P. S. (1990). "Variance estimation when a first phase area sample is restratified." *Survey Methodology* **16**: 99-103.

Kyriakidis, P. C. and J. L. Dungan (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* **8**(4): 311-330.

Lambin, E. and D. Ehrlich (1996). "The surface temperature-vegetation index space for land cover and land cover change analysis." *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.

Lillesand, T. M., R. W. Kiefer and J. W. Chipman (2003). *Remote Sensing and Image Interpretation, Fifth Edition*. New York, John Wiley & Sons, 784 pp.

Little, R. J. A. (2004). "To model or not to model? Competing modes of inference for finite population sampling." *Journal of the American Statistical Association* **99**: 456-556.

Liu, W. G. and E. Y. Wu (2005). "Comparison of non-linear mixture models: Sub-pixel classification." *Remote Sensing of Environment* **94**(2): 145-154.

Lohr, S. L. (1999). *Sampling: Design and Analysis*. New York, Duxbury Press, 450 pp.

Lotsch, A., Y. Tian, M. A. Friedl and R. B. Myneni (2003). "Land cover mapping in support of LAI and FPAR retrievals from EOS-MODIS and MISR: Classification methods and sensitivities to errors." *International Journal of Remote Sensing* **24**(10): 1997-2016.

Loveland, T. R., B. C. Reed, J. F. Brown, D. O. Ohlen, Z. Zhu, L. Yang and J. W. Merchant (2000). "Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data." *International Journal of Remote Sensing* **21**(6-7): 1303-1330.

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

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

Macleod, R. D. and R. G. Congalton (1998). "Quantitative comparison of change-detection algorithms for monitoring eelgrass from remotely-sensed data." *Photogrammetric Engineering and Remote Sensing* **64**: 207-216.

Magnussen, S., S. V. Stehman, P. Corona and M. A. Wulder (2004). "A Polya-urn resampling scheme for estimating precision and confidence intervals under one-stage cluster sampling: Application to map classification accuracy and cover-type frequencies." *Forest Science* **50**(6): 810-822.

Maselli, F., C. Conese and L. Petkov (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.

Mayaux, P. and E. F. Lambin (1995). "Estimation of tropical forest area from coarse spatial-resolution data—a two-step correction function for proportional errors due to spatial aggregation." *Remote Sensing of Environment* **53**(1): 1-15.

McIver, D. K. and M. A. Friedl (2001). "Estimating pixel-scale land cover classification confidence using non-parametric machine learning methods." *IEEE Transactions on Geoscience and Remote Sensing* **39**(9): 1959-1968.

Moody, A. and C. E. Woodcock (1994). "Scale-dependent errors in the estimation of land-cover proportions—Implications for global land-cover datasets." *Photogrammetric Engineering*

and Remote Sensing **60**(5): 585-594.

Morisette, J. T., J. L. Privette and C. O. Justice (2002). "A framework for the validation of MODIS Land products." *Remote Sensing of Environment* **83**(1-2): 77-96.

Morisette, J. T., J. L. Privette, A. Strahler, P. Mayaux and C. O. Justice (2003). "An approach for the validation of global land cover products through the Committee on Earth Observing Satellites." In: *Remote Sensing and GIS Accuracy Assessment*. R. S. Lunetta and J. G. Lyon, Eds. Boca Raton, FL, CRC Press, pp. 304.

Muchoney, D. M. and A. H. Strahler (2002). "Regional vegetation mapping and direct land surface parameterization from remotely sensed and site data." *Int. J. Remote Sens.* **23**(6): 1125-1142.

Myhre, G. and A. Myhre (2003). "Uncertainties in radiative forcing due to surface albedo changes caused by land-use changes." *Journal of Climate* **16**(10): 1511-1524.

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

Naesset, E. (1996*b*). "Conditional tau coefficient for assessment of producer's accuracy of classified remotely sensed data." *ISPRS Journal of Photogrammetry and Remote Sensing* **51**: 91-98.

Nusser, S. M. and J. J. Goebel (1997). "The National Resources Inventory: A long-term multi-resource monitoring programme." *Environmental and Ecological Statistics* **4**: 181-204.

Nusser, S. M. and E. E. Klaas (2003). "Survey methods for assessing land cover map accuracy." *Environmental and Ecological Statistics* **10**: 309-331.

Overton, W. S. and S. V. Stehman (1996). "Desirable design characteristics for long-term monitoring of ecological variables." *Environmental and Ecological Statistics* **3**: 349-361.

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

Powell, R. L., N. Matzke, C. de Souza, M. Clark, I. Numata, L. L. Hess and D. A. Roberts (2004). "Sources of error in accuracy assessment of thematic land-cover maps in the Brazilian Amazon." *Remote Sensing of Environment* **90**: 221-234.

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

Riebsame, W. E., W. J. Parton and K. A. Galvin (1994). "Integrated modeling of land-use and cover change." *BioScience* **44**(5): 350-356.

Rosenfield, G. H., K. Fitzpatrick-Lins and H. S. Ling (1982). "Sampling for thematic map accuracy testing." *Photogrammetric Engineering and Remote Sensing* **48**: 131-137.

Roy, D. P., P. E. Lewis and C. O. Justice (2002). "Burned area mapping using multi-temporal moderate spatial resolution data—a bi-directional reflectance model-based expectation approach." *Remote Sensing of Environment* **83**: 263-286.

Royall, R. M. and K. R. Eberhardt (1975). "Variance estimates for the ratio estimator." *Sankhya* **C 37**: 43-52.

Särndal, C. E., B. Swensson and J. Wretman (1992). *Model-Assisted Survey Sampling*. New York, Springer-Verlag, 694 pp.

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

Schneider, A., M. A. Friedl, D. K. McIver and C. E. Woodcock (2003). "Mapping urban areas by fusing multiple sources of coarse resolution remotely sensed data." *Photogrammetric*

Engineering and Remote Sensing **69**(12): 1377-1386.

Sellers, P. J., C. J. Tucker, G. J. Collatz, S. O. Los, C. O. Justice, D. A. Dazlich and D. A. Randall (1994). "A global 1-degree-by-1-degree NDVI data set for climate studies. 2. The generation of global fields of terrestrial biophysical parameters from the NDVI." *International Journal of Remote Sensing* **15**(17): 3519-3545.

Shalan, M. A., M. K. Arora and J. Elgy (2004). "CASCAM: Crisp and soft classification accuracy measurement software." In: *GeoDynamics: Modelling Spatial Change and Process*. P. M. Atkinson, G. M. Foody, S. E. Darby and F. Wu, Eds. Boca Raton, CRC Press, 440 pp.

Sheppard, C. R. C., K. Matheson, J. C. Bythell, P. Murphy, C. B. Myers and B. Blake (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.

Smith, J. H., S. V. Stehman, J. D. Wickham and L. M. Yang (2003). "Effects of landscape characteristics on land-cover class accuracy." *Remote Sensing of Environment* **84**(3): 342-349.

Smits, P. C., S. G. Dellepiane and R. A. Schowengerdt (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., D. A. Patterson and R. L. Redmond (2003). "Toward estimation of map accuracy without a probability test sample." *Environmental and Ecological Statistics* **10**: 333-356.

Steele, B. M., J. C. Winne and R. L. Redmond (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. (1996). "Use of auxiliary data to improve the precision of estimators of thematic map accuracy." *Remote Sensing of Environment* **58**: 169-176.

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

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

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

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

Stehman, S. V. (2000). "Practical implications of design-based sampling inference for thematic map accuracy assessment." *Remote Sensing 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. (2004*a*). "A critical evaluation of the normalized error matrix in map accuracy assessment." *Photogrammetric Engineering and Remote Sensing* **70**: 743-751.

Stehman, S. V. (2004*b*). "Sampling designs for accuracy assessment of large-area, land-cover maps: Challenges and future directions." In: *Remote Sensing and GIS Accuracy Assessment*. R. S. Lunetta and J. G. Lyon, Eds. New York, CRC Press, pp. 13-29.

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

Stehman, S. V., R. L. Czaplewski, S. M. Nusser, L. Yang and Z. Zhu (2000). "Combining accuracy assessment of land-cover maps with environmental monitoring programs." *Environmental Monitoring and Assessment* **64**: 115-126.

Stehman, S. V., J. D. Wickham, L. Yang and J. H. Smith (2003). "Accuracy of the national land-cover dataset (NLCD) for the eastern United States: Statistical methodology and regional results." *Remote Sensing of Environment* **86**: 500-516.

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

Swain, P. H. (1978). "Fundamentals of pattern recognition in remote sensing." In: *Remote Sensing: The Quantitative Approach*. P. H. Swain and S. M. Davis, Eds. New York, McGraw Hill, pp. 136-187.

Tansey, K., J. M. Gregoire, D. Stroppiana, A. Sousa, J. Silva, J. M. C. Pereira, L. Boschetti, M. Maggi, P. A. Brivio, R. Fraser, S. Flasse, D. Ershov, E. Binaghi, D. Graetz and P. Peduzzi (2004). "Vegetation burning in the year 2000: Global burned area estimates from SPOT VEGETATION data." *Journal of Geophysical Research-Atmospheres.* **109**(D14S03).

Thomas, I. L. and G. M. Allcock (1984). "Determining the confidence level for a classification." *Photogrammetric Engineering and Remote Sensing* **50**: 1491-1496.

Thompson, S. K. (1992). *Sampling*. New York, Wiley, 343 pp.

Tian, Y., R. E. Dickinson, L. Zhou, X. Zeng, Y. Dai, R. B. Myneni, Y. Knyazikhin, X. Zhang, M. Friedl, H. Yu, W. Wu and M. Shaikh (2004). "Comparison of seasonal and spatial variations of leaf area index and fraction of absorbed photosynthetically active radiation from Moderate Resolution Imaging Spectroradiometer (MODIS) and Common Land Model." *Journal of Geophysical Research-Atmospheres* **109**(D1), Art. n° D1103.

Townshend, J., C. Justice, W. Li, C. Gurney and J. McManus (1991). "Global land cover classification by remote sensing—Present capabilities and future possibilities." *Remote Sensing of Environment* **35**(2-3): 243-255.

Townshend, J. R. G., V. Bell, A. Desch, C. Havlicek, C. Justice, W. L. Lawrence, D. Skole, W. Chomentowski, B. Moore III, W. Salas and C. J. Tucker (1995). *The NASA Landsat Pathfinder Humid Tropical Deforestation Project*. Proceedings Land Satellite Information in the Next Decade, ASPRS Conference, Vienna Virginia, 25-28th Sep. 1995, pp. IV-76 - IV-87.

Townshend, J. R. G. and C. O. Justice (2002). "Towards operational monitoring of terrestrial systems by moderate-resolution remote sensing." *Remote Sensing of Environment* **83**(1-2): 351-359.

Trodd, N. M. (1995). *Uncertainty in land cover mapping for modelling land cover change*. Proceedings of RSS 95: Remote Sensing in Action, Nottingham, Remote Sensing Society, pp. 1138-1145.

Turk, G. (2002). "Map evaluation and "chance correction" (letter)." *Photogrammetric Engineering and Remote Sensing* **68**(2): 123-126.

USFS (1992). Forest Service Resource Inventories: An Overview. Forest Inventory, Economics and Recreation Research, USGPO 1992-241-350/60861, Washington, D. C., U.S. Department of Agriculture, Forest Service, 39 pp.

USGS-EDC (2003). "Multi-Resolution Land Characteristics 2001 (MRLC2001)." EROS Data Center, U.S. Geological Survey, http://edc.usgs.gov/products/satellite/mrlc2000.html, 04/24/03.

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

van Oort, P. A. J., A. K. Bregt, S. de Bruin, A. J. W. de Wit and A. Stein (2004). "Spatial variability in classification accuracy of agricultural crops in the Dutch national land-cover database." *International Journal of Geographical Information Science* **18**: 611-626.

Wickham, J. D., S. V. Stehman, J. H. Smith, T. G. Wade and L. Yang (2004). "A priori evaluation of two-stage cluster sampling for accuracy assessment of large-area land-cover maps." *International Journal of Remote Sensing* **25**: 1235-1252.

Wickham, J. D., S. V. Stehman, J. H. Smith and L. Yang (2004). "Thematic accuracy of the 1992 National Land-Cover Data for the western United States." *Remote Sensing of Environment* **91**: 452-468.

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

Zhan, X., R. S. DeFries, J. R. G. Townshend, C. Dimiceli, M. Hansen, C. Huang and R. Sohlberg (2000). "The 250m global land cover change product from the Moderate Resolution Imaging Spectroradiometer of NASA's Earth Observing System." *International Journal of Remote Sensing* **21**(6-7): 1433-1460.

Zhang, X. Y., M. A. Friedl, C. B. Schaaf, A. H. Strahler and A. Schneider (2004). "The footprint of urban climates on vegetation phenology." *Geophysical Research Letters* **31**(12).

Zhou, L., R. E. Dickinson, Y. Tian, X. Zeng, Y. Dai, Z. L. Yang, C. B. Schaaf, F. Gao, Y. Jin, A. Strahler, R. B. Myneni, H. Yu, W. Wu and M. Shaikh (2003). "Comparison of seasonal and spatial variations of albedos from Moderate-Resolution Imaging Spectroradiometer (MODIS) and Common Land Model." *Journal of Geophysical Research-Atmospheres* **108**(D15), Art. n° 4795.