International Journal of Remote Sensing
Vol. 27, No. 4, 20 February 2006, 663–683

Taylor & Francis
Taylor & Francis Group

An accuracy assessment framework for large-area land cover classification products derived from medium-resolution satellite data

MICHAEL A. WULDER*†, STEVEN E. FRANKLIN‡, JOANNE C. WHITE†, JULIA LINKE§ and STEEN MAGNUSSEN†

†Canadian Forest Service, 506 West Burnside Road, Victoria, BC, Canada V8Z 1M5
‡Department of Geography, University of Saskatchewan, Saskatoon, Saskatchewan, Canada S7N 5A5
§Department of Geography, University of Calgary, Calgary, Alberta, Canada T2N 1N4

(*Received 28 February 2005; in final form 20 April 2005*)

Land cover classification over large geographic areas using remotely sensed data is increasingly common as a result of the requirements of national inventory and monitoring programmes, scientific modelling and international environmental treaties. Although large-area land cover products are more prevalent, standard operational protocols for their validation do not exist. This paper provides a framework for the accuracy assessment of large-area land cover products and synthesizes some of the key decision points in the design and implementation of an accuracy assessment from the literature. The fundamental components of a validation plan are addressed and the framework is then applied to the land cover map of the forested area of Canada that is currently being produced by the Earth Observation for Sustainable Development programme. This example demonstrates the compromise between the theoretical aspects of accuracy assessment and the practical realities of implementation, over a specific jurisdiction. The framework presented in this paper provides an example for others embarking on the assessment of large-area land cover products and can serve as the foundation for planning a statistically robust validation.

1. Introduction

The classification of land cover over large geographic areas with remotely sensed data is increasingly common. An overview of the status and research priorities for large-area mapping with satellites is given by Cihlar (2000). Regions (Homer *et al.* 1997), nations (Loveland *et al.* 1991, Fuller *et al.* 1994, Zhu and Evans 1994, Cihlar and Beaubien 1998), continents (Stone *et al.* 1994) and the globe (Loveland and Belward 1997, Loveland *et al.* 2000, Hansen *et al.* 2000, Zhu and Waller 2003) have been mapped with a range of satellite data at various spatial resolutions. The increase in large-area land cover mapping projects may be explained by several factors: an increase in the availability and/or affordability of low spatial resolution (AVHRR, MODIS) and medium spatial resolution (Landsat, SPOT, IRS) imagery; a need for national and global land cover inventories to facilitate the modelling of complex Earth systems; the study of land cover change resulting from deforestation and environmental degradation; and the political need to fulfil obligations of

*Corresponding author. Email: mwulder@nrcan.gc.ca

International Journal of Remote Sensing
ISSN 0143-1161 print/ISSN 1366-5901 online
http://www.tandf.co.uk/journals
DOI: 10.1080/01431160500185284

international treaties such as the Convention on Climate Change (Kyoto Protocol) and the Convention on Biological Diversity.

Although large-area land cover products are increasingly prevalent, standard operational protocols for their validation do not exist (Justice *et al.* 2000). Furthermore, a sufficient level of accuracy is assumed in order to rationalize the applied use of these products (Morisette *et al.* 2002, Stehman and Czaplewski 2003). Given the significance of these large-area land cover products to a wide variety of applications, their validation in a statistically robust fashion is vital. As accuracy assessment protocols for large-area land cover products are developed, some of the issues that need to be addressed include: logistically feasible and statistically valid sampling strategies; the capability to assess the accuracy of the reference data; and stable and informative metrics of accuracy (Franklin and Wulder 2002).

The validation of map products derived from remotely sensed data has evolved significantly over time. Congalton (1994) categorized four definite historical stages of accuracy assessment, progressing from visual appraisal, to aspatial area summaries, to primitive error matrix analysis, to the current rigorous statistical approaches. There have been several reviews of accuracy assessment methods in the literature over the past 15 years (Congalton 1991, Janssen and van der Wel 1994, Justice *et al.* 2000, Foody 2002). Over this time, accuracy assessment protocols have changed in response to the demands of statistical rigour and the increasingly sophisticated applications that are now possible with satellite-based land cover maps (e.g. Bauer *et al.* 1994, Cohen *et al.* 2001). The site-specific methods used to assess accuracy at local and regional scales may not be fully transferable to coarser (i.e. national and global) scales; however, several key elements of existing remote sensing validation methodologies form the foundation for the design of an accuracy assessment framework for large-area land cover products.

The objective of this paper is to provide a framework for the accuracy assessment of large-area land cover products and to develop some context and understanding for the application of the framework to a specific product: the Earth Observation for Sustainable Development (EOSD) land cover of the forested area of Canada (Wulder *et al.* 2003). Large-area land cover products require well thought out validation plans that have clearly defined goals and objectives and that identify and execute proper statistical procedures. In this paper, we present a summary of the principles and techniques appropriate for the accuracy assessment of large-area land cover classifications developed from satellite remotely sensed data. First, we consider several elements of existing accuracy assessment protocols, relying on the literature to reveal key decision-points that must be considered in the validation of a large-area land cover mapping product. Then a discussion of accuracy assessment, within the context of the EOSD mandate, is provided. The principles developed and presented may be adapted to suit the circumstances of other jurisdictions facing a similar need to assess the accuracy of a large-area land cover product.

2. Establishing a framework large-area land cover accuracy assessment

A framework facilitates a standardized decision-making process whereby key issues in the design of an accuracy assessment for a large-area land cover product are addressed. Initially, it is imperative that the objective of the accuracy assessment be clearly defined, followed by a thorough consideration of the three essential components of an accuracy assessment: sampling design, response design and analysis (Stehman and Czaplewski 1998).

2.1 *Special considerations for large-area land cover*

Statistically rigorous accuracy assessment for large-area land cover classification is frequently constrained by the lack of suitable ground data, logistical realities and high monetary costs (Merchant *et al*. 1994, Muchoney *et al*. 1999). A decade ago, Congalton (1994) identified the need for procedures to validate large-area products. More recently, Foody (2002) highlighted many of the issues associated with the accuracy assessment of large-area mapping products derived from coarse-resolution remotely sensed data. The primary methodological error cited by Foody (2002) is the reporting of invalid accuracy assessment results derived from standard error matrix analysis, despite the violation of many of the assumptions associated with this type of analysis (e.g. pure pixels, discrete classes). There is growing acceptance that traditional means of validating classifications generated from Landsat or SPOT data are not automatically transferable to coarser scales (e.g. AVHRR) (Merchant *et al*. 1994, Foody 2002).

It is important to make the distinction between the validation of large-area land cover products derived from low spatial resolution (AVHRR, MODIS) imagery and those products derived from medium spatial resolution (Landsat, SPOT) imagery, as each product presents unique challenges for validation (Merchant *et al*. 1994, Scepan 1999, Wulder *et al*. 2004a). However, all large-area land cover products, regardless of the spatial resolution of the source data from which they are generated, share many common logistical constraints that primarily result from the large spatial extent of the products.

This paper focuses on the validation of products derived from medium spatial resolution data (Franklin and Wulder 2002). Examples of these products include national land cover maps derived from Landsat or SPOT data, such as the large-area land cover product of the conterminous United States produced by the Multi-Resolution Land Characteristics Consortium (Vogelmann *et al*. 2001) and the EOSD land cover mapping project of the Canadian Forest Service and the Canadian Space Agency (Wulder *et al*. 2003). Both of these products present challenges for validation (Stehman 1996b). However, as more large-area land cover products are generated and validated, appropriate and robust sampling and analysis protocols are emerging (Edwards *et al*. 1998, Scepan 1999, Zhu *et al*. 1999, Stehman *et al*. 2000, Yang *et al*. 2001, Morisette *et al*. 2002, Fuller *et al*. 2003, Stehman *et al*. 2003).

2.2 *Defining the objective of the accuracy assessment*

Map accuracy is a sufficiently complex concept that no single definition or approach can realistically encompass all possible situations in answering the question: How accurate is the map? The appropriate accuracy assessment protocol often develops from consideration of the following question: Is the map sufficiently accurate for a specific application? The understanding is that not all applications require the same level of accuracy to be successfully accomplished, and therefore the same level of effort need not be expended to determine map accuracy for different possible applications. Unfortunately, in the context of large-area land cover products, the needs of all possible future applications cannot be foreseen or accommodated (Stehman 1996b).

Accuracy assessment may be undertaken for many different purposes, and these different purposes may influence the choice of approach and result (Janssen and van der Wel 1994). For some users, the lack of an appropriately detailed accuracy

assessment would be cause to discontinue the use of the map. For others, the absence of an accuracy assessment might constitute little or no impediment to map acceptance and use; Stehman (2001) suggested that if probability-designed sampling cannot be carried out, or if sample sizes are too small such that estimates have poor precision, it would be prudent to omit the assessment entirely — since even minimal standards of rigour would not be met. Similarly, Czaplewski (2003) identified the minimal value gained by investing scarce resources in an unreliable accuracy assessment. The idea of forgoing an accuracy assessment under certain conditions requires consideration in juxtaposition to the substantial research effort in the field of remote sensing that is focused on methods for increasing accuracy. If a map is expected to be less than adequate for a given application, perhaps resources are best expended on improving the classification accuracy to a desired level, as opposed to documenting map accuracy deficiencies in comprehensive detail.

The objective of the accuracy assessment will ultimately determine the scope of the assessment, the degree of statistical rigour necessary, and the requisite level of resources. Clearly defining the objective of the accuracy assessment in consideration of the intended application(s) and the operational constraints of the specific project determines whether an accuracy assessment is a worthwhile investment. An unambiguous objective sets the stage for the design of the accuracy assessment. The rationale for the stated objective should be well documented, to ensure that future users of the land cover product clearly understand the circumstances within which the accuracy statistics may be appropriately applied. In the context of statistical estimation, accuracy and precision are closely associated with one another. Accuracy defines how close the sample map accuracy is to the true map accuracy, while precision defines the repeatability of obtaining the sample map accuracy (or a value close to it) through repetitive sampling. An estimate with high precision in this context refers to one with low variability and a narrow confidence interval (Stehman 2001). Both accuracy and precision should be documented.

3. Sample design

The sample design is the protocol by which the reference data are selected (Stehman and Czaplewski 1998). A sample design is required to determine the number and spatial location of samples that will be obtained for the accuracy assessment. Inherent to the process of selecting an appropriate sampling design is the selection of the sampling frame and sampling unit.

3.1 *Determining the appropriate sampling frame*

A sampling frame is defined as the 'materials or devices, which delimit, identify, and allow access to the elements of the target population' (Särndal *et al.* 1992, p. 9). Ideally, the sampling frame will match the target population; however, in natural resource applications it is common for the sampling frame to contain many non-target elements, and for many of the target elements to be inaccessible (Stevens and Olsen 2004). Sampling frames are usually list frames or area frames; in an ecological context, both forms of sampling frames may be implemented, depending on the nature of the sample units (Stevens 1994). For example, discrete features, such as small lakes, may be represented as points even though they have some area associated with them. Conversely, extensive features, such as forests, may be represented in an area frame. A list frame, as the name implies, is a list of all the

sample units available for sampling. The type of sampling frame that is most appropriate for large-area land cover validation depends on the nature of each land cover class and whether the class is discrete or extensive. When using an area frame, an explicit rule for associating a spatial location within the frame to a unique sampling unit must be established. List frames may also be spatially enabled. A sample frame may be defined by a specific spatial dataset, such as the National Air Photo Program coverage (Zhu *et al*. 2000) or the entire landmass of a country (if evaluating a national product).

3.2 *Determining the appropriate sampling unit*

Sampling units are usually either point or area and their various forms and relative advantages and disadvantages are detailed in Stehman and Czaplewski (1998) and Stehman and Overton (1996). In some circumstances, discrete areal units such as lakes may be represented as points for sampling purposes (Stevens 1994, Stevens and Olsen 2004). Commonly, area units such as pixels, polygons or fixed-area plots are used (e.g. Zhu *et al*. 2000, Laba *et al*. 2002, Stehman *et al*. 2003). Note that the choice of sampling unit is not constrained by the unit of representation on the map, and that the sampling unit is identified independent of the reference classification (Stehman and Czaplewski 1998). Selection of an appropriate sampling unit may be influenced by: project objectives, landscape characteristics, features of the mapping process, practical constraints, location error, minimum mapping unit, and treatment of polygon boundaries (Stehman and Czaplewski 1998, Zhu *et al*. 2000). In the context of large-area land cover, some have selected the sampling unit to reflect the unit of the final product (Stehman 1996a, Zhu *et al*. 1999) or the unit of the source imagery (Franklin *et al*. 1991, Janssen and van der Wel 1994).

3.3 *Selecting an appropriate sample size*

Sample size is often determined based on probability theory (binomial distribution) or an approximation of the normal distribution (Congalton 2001). There is a clear relationship between sample size and total population size that suggests a threshold beyond which additional samples add little to the assessment of accuracy (although increasing confidence may be obtained). For an accuracy assessment of a land cover classification, a sample size of 100 per class ensures that accuracy can be estimated with a standard error of no greater than 0.05 (Stehman 2001). Many authors recommend 30 to 50 samples per class (e.g. Goodchild *et al*. 1994). Smaller samples represent lower cost as well as lower precision; if land cover classes have relative levels of importance to the objective of the accuracy assessment, then reducing the sample size in less important classes can preserve precision for the more important classes (Stehman 2001). In image data, location-dependent models of uncertainty and spatial autocorrelation may need to be considered since these properties of spatial data can influence sample sizes useful for a variety of image processing functions, including accuracy assessment (Moisen *et al*. 1994, Kyriakidis and Dungan 2001). Czaplewski (2003) recommends the construction of an expected error matrix in order to determine the appropriate sample size for any given application. This expected error matrix facilitates the determination of confidence intervals for alternate scenarios of sample design and sample size, allowing the user to evaluate tradeoffs between the costs of sampling and the desired precision of accuracy estimates.

3.4 *Selecting an appropriate target accuracy*

A fundamental question of sample design involves determining the acceptable level of accuracy, since the acceptable or target accuracy, combined with the acceptable level of error, may be used to determine the required sample size. The question of the acceptable level of accuracy is often answered by reference to the seminal work of Anderson *et al*. (1976) who outline the criteria for an effective land use and land cover classification scheme for use in conjunction with remotely sensed data. Specifically, Anderson *et al*. (1976, p. 5), citing the earlier work of Anderson (1971), state that 'the minimum level of interpretation accuracy in the identification of land use and land cover categories from remote sensor data should be at least 85 percent'. A review of the original source indicates that Anderson (1971, p. 381) recommended 'a minimum level of accuracy of about 85 to 90 percent or better'. Anderson (1971, p. 381) rationalizes this level of accuracy as being 'nearly comparable with the level of accuracy attained by the Bureau of the Census in obtaining information about land use by enumeration in the Census of Agriculture, which is taken every five years'. Therefore, although an 85% accuracy target is widely accepted by the remote sensing community as a benchmark, as several recent examples indicate (Foody 2002, Reese *et al*. 2002, Fuller *et al*. 2003, Tømmervik *et al*. 2003), its usefulness as a standard is unclear. Others have also questioned the validity of the 85% target (Laba *et al*. 2002, p. 453):

> The accuracy assessments of several recently completed regional-scale land cover mapping projects indicate that producer's and user's accuracies are stabilizing in the 50–70% range, independent of level of taxonomic detail or methodological approaches (Edwards *et al*. 1998, Ma *et al*. 2001, Zhu *et al*. 2000). In our view, additional improvements in accuracy are not likely, and that only through the use of sensors with high spectral, spatial, and temporal resolution will map accuracies approach 80%. In addition, the traditional, and artificial, target of 85% overall percent correct should not be used as a criterion to measure success or failure of a land cover mapping project.

Is 85% a realistic goal for all classes or only a subset? Are vegetation classes given equal importance (or weight) with non-vegetated classes and are area weights appropriate as per Card (1982)? Anderson *et al*. (1976) suggested that the accuracy for all classes should be about equal; such an approach might imply that an effort to obtain consistent accuracies, at some lower level, may be a priority over achieving higher accuracy in only a few classes. Given a national programme of land cover mapping with highly variable input data layers and resources for image analysis, there may be some rationale to examine carefully the viability of an 85% accuracy target and the ideas of approximately equal accuracy and weights across all classes.

Ultimately, the required level of accuracy of the product should depend on its intended application. For example, Czaplewski and Patterson (2003) describe a remotely sensed land cover product that was created for the purpose of stratifying a landscape in order to improve sampling efficiency. Czaplewski and Patterson (2003) provide a summary of the accuracies required for predetermined gains in sampling efficiency, as a function of the number of categories in the classification. For example, in order to garner substantial gains in sampling efficiency from a land cover product having more than 10 classes, accuracies for each stratum are required to exceed 70%. Czaplewski and Patterson (1993) identify a strong dependence between accuracy and prevalence. If a subpopulation is common, then the

classification accuracy must be very high; conversely, if the subpopulation were rare then a lower level of accuracy would suffice (for this application).

3.5 *Approaches for evaluating map accuracy*

The key uncertainties in developing an accuracy assessment protocol flow from the tradeoffs between samples collected (n), p value (error bar widths) and the choice or selection of the description of accuracy. Czaplewski (2003) and Goodchild *et al.* (1994) provided a comprehensive discussion of these issues. Goodchild *et al.* (1994) outlined three possible scenarios for evaluating map accuracy. The first, which was originally described by Aronoff (1985), involves identifying the minimum map accuracy that would cause the null hypothesis (associated with a specified target accuracy) to be rejected. The second determines, with a required level of certainty, that the desired level of accuracy has been achieved. The third approach assesses, for a given sample map accuracy, the probability that the actual map accuracy equals or exceeds a specified target.

The first approach described above is perhaps the most suitable for large-area land cover product validation because it facilitates the use of smaller sample sizes. A one-sided z-test statistic can be used to identify the range of minimum map accuracies that would not cause rejection of the null hypothesis (e.g. where $H_0=0.80$). The z-test statistic may be determined as follows:

$$z = (\hat{p} - p) \Big/ \sqrt{p(1-p)/n} \tag{1}$$

where $\hat{p}$ is the sample map accuracy, p is the target map accuracy and n is the size of the sample. If the calculated z value is greater than -2.36 (99% confidence level), -1.65 (95% confidence level), or -1.29 (90% confidence level) then the null hypothesis is not rejected ($H_0 : p = \hat{p}$). Note that equation (1) is based on the binomial variance of binary sample data, and is therefore valid only for simple random sampling of individual pixels with a binary attribute (i.e. correct, incorrect). If other sampling strategies are used, such as stratified or cluster sampling, a correction for the difference in design (design effect) must be applied to the sample size (Kish 1967). Table 1 contains some examples of minimum accuracies for a small number of sample sizes, at various significance levels, calculated using equation (1). Czaplewski (2003) provides similar tables to aid in the selection of an appropriate sample size based on desired levels of accuracy. These tables help the analyst evaluate the costs and benefits associated with increasing sample size and thereby

Table 1. Minimum acceptable map accuracies ($\hat{p}$ values) by significance levels and sample size, assuming a target accuracy of 80%.

	Significance level (α)		
Sample size (n)	0.01	0.05	0.10
20	0.589	0.652	0.685
100	0.701	0.734	0.748
300	0.745	0.761	0.770
500	0.757	0.770	0.777
1000	0.770	0.779	0.783

Adapted from Goodchild *et al.* (1994).

increase the precision of the estimates. Clearly, as sample size increases, the confidence with which one could reject the null hypothesis increases (Aronoff 1985). The minimum accuracy approach is very tolerant, as a sample map accuracy must be low before the assertion that the sample accuracy differs from the target accuracy may be rejected.

Another form of the minimum accuracy approach is to determine the sample size appropriate for a given acceptable error (deviation from target accuracy) and significance level:

$$n = \frac{pq}{[E/z_\alpha]^2} \tag{2}$$

where p is the required accuracy of the data, q is $(1-p)$, E is the acceptable error and z_α is the α-quantile of the normal. Table 2 is a compilation of the appropriate sample sizes for a given level of acceptable error and significance. In order to account for the likelihood that not all of the selected samples can be validated (for logistical or other reasons), sample size is often increased to ensure that the minimum sample size is attainable (Nusser and Klaas 2003, Stehman and Czaplewski 2003).

3.6 *Selecting the sample design*

At this point in establishing an accuracy assessment framework, an objective for the validation has been clearly defined, the sample frame has been determined, the sampling unit has been selected, an appropriate target accuracy has been chosen and a sample size that facilitates hypothesis testing relative to that target accuracy has been identified. Now the question of where to sample must be addressed by the selection of an appropriate sample design. The selection of the sample design is perhaps the most complex decision to be made in the planning of an accuracy assessment. Several sources provide reliable guidance on the selection of sample design (Skidmore and Turner 1992, Moisen *et al*. 1994, Stehman and Overton 1996, Stehman 1996a, b, 1999, Kalkhan 1998, Stehman and Czaplewski 1998). Stehman and Czaplewski (2003) note that there are two specific design criteria that are desirable for large-area land cover validation: cluster sampling to reduce travel costs for field visitation or to reduce acquisition costs for aerial photography, and stratification by land cover class, to enable accuracy parameters to be reported by class with adequate levels of precision. Examples of accuracy assessments which incorporate both of these design features are Yang *et al*. (2001) and Nusser and Klaas (2003).

Table 2. Sample sizes versus acceptable error.

Level of acceptable error	80% accuracy		
	90% CI (n)	95% CI (n)	99% CI (n)
0.01	2663	4356	8911
0.03	296	484	990
0.05	107	174	356
0.10	27	44	89
0.20	7	11	22

Adapted from Goodchild *et al*. (1994).
n, number of Samples.

Czaplewski (2003) noted that simple sample designs are preferred as they preclude reliance on professional statisticians and allow those who produce the maps to conduct the accuracy assessment directly; however, the complexity of a large-area land cover validation exercise necessitates the expertise of a professional statistician. As outlined by Stehman *et al.* (2003), a sampling design should comply with four criteria: standards of defining a probability sample must be adhered to; adequate sample sizes must be used for estimating user's accuracy with acceptable levels of precision; cost efficiencies must be considered; and the spatial distribution of samples must be representative across the area of interest. The use of a probability sampling design, where each sampling unit has a known likelihood of inclusion in the sample, is critical for statistical validity and analysis (Czaplewski 2003). Cost efficiency may be achieved by the selection of a sampling strategy 'that can provide the same variance using fewer sampling locations than another strategy' (Stehman 1996b, p. 486). For large-area land cover product validation, Stehman (1996b) suggested that stratification by large geographic regions would be useful to ensure a spatially representative sample of adequate size to support regional reporting. Stratification by land cover class places a higher priority on the calculation of user's accuracy and commission errors (Aronoff 1982, Stehman 1995). Sample size is linked to the sample design; therefore, sample size should be revisited after a sample design has been selected (Czaplewski and Patterson 2003).

4. Response design

The response design is the protocol for determining the reference classification recorded at each sampling unit (Stehman 2001). There are two distinct components of a response design: the evaluation protocol, which are the procedures used to collect the reference information, and the labelling protocol, which specifies the land cover label that will be assigned to the sampling unit.

4.1 *Selecting an evaluation protocol*

In defining the evaluation protocol, the spatial support region (SSR) is chosen. The SSR is defined as 'the size, geometry, and orientation of the space on which an observation is defined' (Atkinson and Curran 1995, p. 768). The SSR is an area surrounding a sampling unit that is used as context for determining the reference classification of the sampling unit. The size and shape of the SSR may vary according the land cover class (e.g. discrete features such as lakes versus extensive features such as forests) (Stevens 1994, Stehman and Czaplewski 1998), may be consistent for all features (Nusser and Klaas 2002, Stehman *et al.* 2003) or may just be the sample unit itself. The impact of conservative bias on accuracy results, caused by spatial registration error and the spatial heterogeneity of the land cover, can be lessened by the wise use of a spatial support region (Verbyla and Hammond 1995). Foody (2002) emphasized the need to incorporate positional tolerance into thematic accuracy assessments, citing the propensity of remote sensing studies to adopt site-specific accuracy assessment procedures when the ability to co-locate sites in the derived map and ground data set is questionable.

The evaluation protocol also includes the methods by which the SSR is used to define the reference classification. For example, a pixel sampling unit could have a 3×3 pixel neighbourhood SSR and the mode of the SSR could be used to determine the reference classification for the sample unit. However, the accuracy estimates

generated are then only applicable to a product that is similarly generalized — not to the original product (Czaplewski, 2003). It is also possible to subsample within the SSR to determine the reference classification (e.g. by using line transects, point samples or cluster plots). This sampling serves only to inform the labelling protocol and is separate from the sample design; however, it may also provide information on the heterogeneity within or surrounding the sample unit (Stehman and Czaplewski 1998).

4.2 *Selecting a labelling protocol*

Based on the information provided by the evaluation protocol, the labelling protocol determines the reference classification of the sampling unit. Labelling is generally restricted to one land cover class; however, some studies have opted to record both a primary and secondary land cover class (Edwards *et al*. 1998, Stehman *et al*. 2003). Finally, the labelling protocol defines the nature of agreement between the map source and the reference source. In Stehman *et al*. (2003) agreement was defined as a match between the mode map class of a 3×3 pixel SSR and either the primary or the secondary label of the reference classification. Note that a SSR can be used to determine the reference classification as well as the rules of agreement between the map and the reference data.

With regard to the manner in which the reference data are classified, Congalton (2001) suggested the reference data must be classified using the exact same classification scheme as the map data in order to reduce confusion. However, a scenario where the reference data has a more detailed classification schema than the map data may be advantageous for statistical estimation (Czaplewski 2003). Unless the reference data are purpose-acquired for the accuracy assessment, they will rarely have the same classification scheme as the map data. The selection of an appropriate source of reference data for large area land cover validation is difficult. Merchant *et al*. (1994: 69) suggested the use of 'a cumulative process of validation that involved consideration of a number of different types of evidence supporting or refuting the accuracy of a given product'. By their very nature, large-area land cover products are spatially extensive and it is unlikely that there will be one source of appropriate reference data that covers the entire area of interest. Logistical constraints often prevent purpose-acquired field measurements, and alternative data sets are frequently used. Examples of reference data sources that are suitable for large-area land cover validation include alternative GIS data sets, purpose-acquired field data, aerial photograph interpretation, and purpose-acquired airborne data.

5. Analysis protocol

The final component of the accuracy assessment is the application of analysis and estimation protocols to the reference sample (Stehman and Czaplewski 1998). Comparisons of the map and reference data are made using a simple cross-tabulation, the product of which is known as a contingency table or a confusion matrix. The confusion matrix is currently at the core of the accuracy assessment literature (Foody 2002), partly because of the many measures of classification accuracy that may be derived from it. Stehman and Czaplewski (1998) suggest that the error matrix should be reported in terms of proportions rather than counts and that an error matrix should never be normalized; rather, preference is given to the use of conditional probabilities based on marginal proportions.

In general, simple accuracy parameters such as the overall classification accuracy and producer's and user's accuracies are easily computed and understood, and are thought to capture the essential quality of the map product (Congalton and Green 1999). Other statistics, such as the kappa coefficient, the errors of omission and commission, and additional summary measures may also be reported; however, the use of a single summarized measures must be treated with caution (Stehman 1997). Confidence levels, and spatial depictions of uncertainty are possible to support the interpretation of accuracy and map quality (McGwire and Fisher 2001).

Estimating accuracy parameters requires incorporation of the inclusion probabilities for the design used (Stehman and Czaplewski 1998). Known inclusion probabilities are required for consistent estimation of accuracy parameters (Stehman 2001), hence the significance of using a probability-based sampling design. Parameter estimation must be done using the correct estimation formula for the selected sample design. Similarly, variance estimation must also be done using the appropriate formula for the given sample design (Czaplewski 1994, 1998, Stehman 1995). The estimated accuracy parameters must be presented with their standard errors to provide the end-user with sufficient information to decide whether they will use the data for their own applications. Finally, the interpretation of the results, in the context of the objectives of the accuracy assessment, is also worthwhile in guiding end-users to appropriate usage of the large-area land cover product.

6. An example large-area land cover accuracy assessment framework

A land cover map of the forested area of Canada is being produced to year 2000 conditions and is scheduled for completion in early 2006 (Wulder *et al*. 2003). Over 400 Landsat scenes will be required to map the forested area of Canada (Wulder and Seemann 2001) with 21 different land cover classes (Wulder and Nelson 2001). Due to the large area involved, differing acquisition seasons and years are being included in the source imagery for the classifications (Wulder *et al*. 2004b). The actual classification is being undertaken by federal, provincial and territorial agencies and private industry. Agencies primarily concerned with the non-forest land base will be mapping other regions, such as agricultural areas and the far north, to represent circa 2000 conditions.

The purpose of this section is to describe how the accuracy assessment framework for large-area land cover could be applied to the EOSD mapping project. The presentation of this framework is intended to serve as an example from which similar procedures for differing jurisdictions may be developed. The primary objective of an assessment of the accuracy of the EOSD product is to report, on a national scale, the overall accuracy of the EOSD large-area land cover product. The secondary objective is to report users' accuracies for individual classes to further facilitate user confidence. In addition, the ability to report overall accuracy by ecozone is also desirable. The emphasis in developing this product is on the forested classes; therefore non-forest classes may be given less weight in the sample design (and therefore will have a lower probability of selection). Accordingly, any tradeoffs requiring a reduction in sample size will be made to non-forest classes. Applied uses of this product include the National Forest Inventory (NFI), biomass estimation, carbon budget modelling, and support for estimation of forest change over time (Wulder *et al*. 2004c). All these applications require information regarding the forested area of Canada, so the objectives of the accuracy assessment, as stated above, correspond with the identified user needs.

6.1 *Sample design*

The sampling frame for this assessment is restricted to the forested area of Canada, which is mapped under the mandate of the EOSD project, as shown in figure 1. The sampling plan is hierarchical, with two levels of stratification and two stages of sample selection (figure 2). This sampling approach is cost-effective and ensures sufficiently large sample sizes for each land cover class stratum (Wickham *et al.* 2004b). An 80% level of accuracy provides an arbitrary benchmark for assessing the EOSD product; however, an alternative accuracy standard may be identified after the completion of an initial pilot study. Such a standard may be based on the level of accuracy required for the intended application of the EOSD product (Czaplewski and Patterson 2003). When

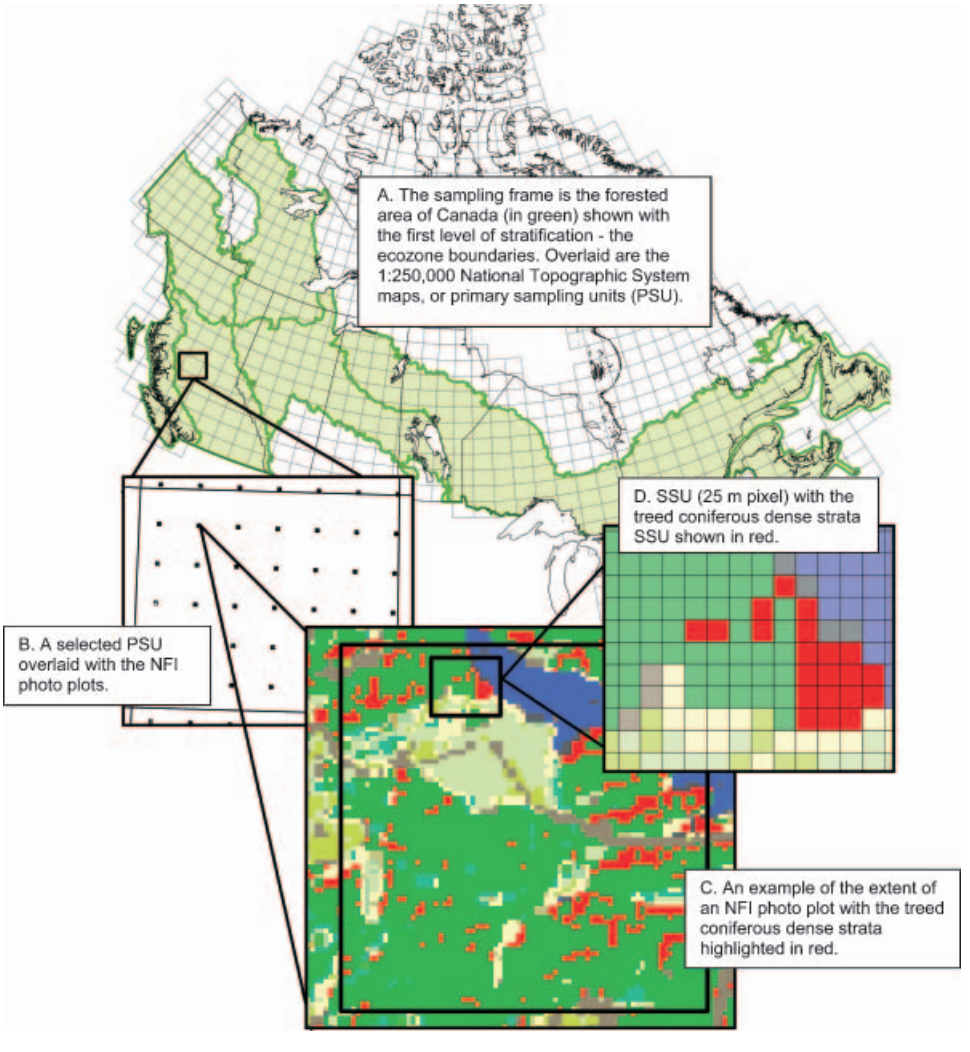

Figure 1. Proposed sample design for EOSD accuracy assessment. (a) The full sampling frame, comprised of the forested area of Canada (green) with the first level of stratification — the ecozone boundaries. Overlaid are the 1 : 250 000 NTS mapsheets or PSU. (b) A sample 1 : 250 000 NTS mapsheet, overlaid with the NFI photo plots. (c) The extent of a NFI photo plot, with the treed coniferous dense strata shown in red. (d) The SSU with the treed coniferous dense strata shown in red.

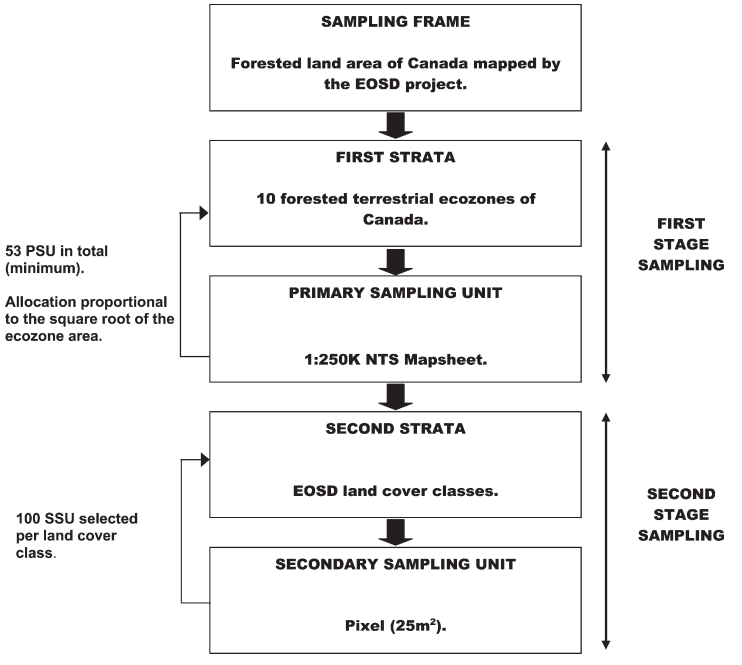

Figure 2. Sampling design for accuracy assessment of EOSD land cover product.

the selected target accuracy is combined with the level of acceptable error (e.g. 10%), an appropriate sample size for the assessment may be calculated.

6.1.1 Allocation of primary sampling units (PSU). For the EOSD sample design, the first level of stratification is formed by the 10 forested ecozones of Canada (figure 1). The ecozones are broad areas of similar ecological characteristics and their use for stratification ensures that the sample is spatially distributed across the country. In addition, the ecozones provide a framework for subsequent investigations into the spatial variability of the EOSD product accuracy. The primary sampling unit (PSU) is the 1 : 250 000 National Topographic System (NTS) mapsheet, which is the unit of product delivery for the EOSD project. Each 1 : 250 000 NTS map sheet is approximately 145 km × 111 km in size (figure 1).

The ecozones vary in size and therefore the number of PSUs within each ecozone also varies (table 3). If PSU were allocated equally to each ecozone, the selection probabilities for the PSU would not be the same in every ecozone. As a result, the variation between the sampling weights would be large, resulting in a large variance in the national estimate of EOSD product accuracy. If the PSU were allocated proportional to the ecozone area, the selection probabilities for all PSU would be equal; unfortunately several of the ecozones would have PSU sample sizes that would make any reporting of accuracy, by ecozone, unreliable. A compromise between equal and proportional allocation of PSU is therefore necessary (Kish 1988). Such a compromise seeks to find a balance between reliable ecozone estimates and a reliable national estimate of accuracy, and may be achieved by allocating the PSU proportional to the square root of the ecozone area (table 3).

6.1.2 PSU sample size determination. Given the parameters of 80% accuracy with a 10% acceptable error, a minimum of 27 PSUs are required to determine whether

Table 3. Allocation of PSU by ecozone.

Ecozone	Number of PSU in ecozone	Area (ha)	Allocation of PSU proportional to ecozone area (no. of PSU selected)	Allocation of PSU proportional to square root of ecozone area (no. of PSU selected)
Taiga Plains	88	62, 025, 662	5	5
Taiga Shield	167	138, 723, 503	11	8
Boreal Shield	196	205, 562, 705	16	10
Atlantic Maritime	35	28, 816, 897	2	4
Boreal Plains	83	71, 213, 468	5	6
Taiga Cordillera	44	25, 124, 722	2	3
Boreal Cordillera	61	44, 476, 028	3	5
Pacific Maritime	40	23, 899, 364	2	3
Montane Cordillera	51	47, 810, 200	4	5
Hudson Plains	50	43, 489, 543	3	4
TOTAL	815	691, 142, 093	53	53

the sampled map accuracy is significantly different from the target of 80% accuracy at the 90% confidence level (see table 2). This sample size is then inflated by 15% to account for non-response, to a total sample size of 31 PSUs. The sample sizes included in table 2 have been derived using the binomial distribution, and an assumption of simple random sample design has been made. As the binomial distribution tends to be overly optimistic in the determination of sample size, and because a two-stage cluster sampling method is being used, the sample sizes will need to be adjusted from the figures presented in table 2. In addition, the impact of spatial autocorrelation of the secondary sampling units (SSUs) on the among-PSU variance must be accounted for (Cochran 1977, Magnussen 2001, Wickham *et al.* 2004a). Due to the size of the PSU (145 km × 111 km) in this study and the number of SSUs each PSU contains (approximately 25.5 million) the likelihood of spatial autocorrelation is high (Campbell 1981). In order to ensure that a sufficient sample size has been selected, the PSU sample size should be increased accordingly.

Magnussen (2001) outlines a method for adjusting sample size to account for the effects of spatial autocorrelation. Based on past research using similar data types in Canada, land cover pixels have been found to be autocorrelated in space by a first-order autoregressive correlation of 0.1 to 0.2. The asymptotic variance inflation due to clustering of SSUs when SSU size goes to infinity is between 2.5 and 3.1. This information can then be used to select a larger sample size, in order to compensate for the impact of spatial autocorrelation on the sample variance. By using the square root of 2.8 (the middle of 2.5 to 3.1 range), a prudent value for sample size inflation is derived (1.7). In our EOSD example, the sample size should therefore be increased from 31 to 53 PSUs.

6.1.3 Selection of secondary sampling units (SSUs). Each selected PSU is stratified by the EOSD land cover classes. Within each land cover class stratum, a random sample of SSUs is selected. The SSU is the 25 m pixel that forms the minimum mapping unit of the EOSD product. A sample of 100 SSUs per land cover class stratum is desirable (Stehman 2001). For example, if a PSU contains 10 different land cover classes, the total number of SSUs will be 1000 (10 classes, 100 samples per

class). Across the country, this could result in a total minimum sample size of 53 000 SSUs (53 PSUs, 10 class per PSU, 100 samples per class), although the final total will depend on the number of different land cover classes found within each PSU. If a PSU is found on an ecozone boundary and the selected SSU is located outside the ecozone, the SSU would not be selected. This maintains the equal probability feature of the second stage of the design.

6.2 *Response design*

One possible reference data source for validation could be the independent GIS and field data available through the National Forest Inventory (NFI) programme (Gillis 2001). The NFI is a new forest inventory programme that will cover all of Canada. In order to provide reliable statistics, the NFI will sample a minimum of 1% of Canada's landmass. This 1% sample translates into a nominal design of 2 km × 2 km plots located on a 20 km × 20 km network, resulting in approximately 20 000 sample plots for Canada. The 2 km × 2 km plots will be identified on conventional, mid-scale aerial photography, and will be delineated and interpreted according to land cover classes and other forest stand attributes. The reference classification between the EOSD product and the NFI are compatible, as the NFI uses the same classification system as the EOSD (Wulder and Nelson 2001). The distribution of the NFI photo plots within any given PSU, as shown in figure 1, does not provide a continuous representation of land cover within the PSU. On average, each PSU will contain approximately 33 NFI photo plots. Issues related to the comparison of raster (EOSD) and vector (NFI) data require further investigation before such a use of the NFI may be committed to.

Where NFI data are not available (e.g. in far northern areas), purpose-acquired airborne video could be a source of reference data. This reference data source will be point samples and, therefore, the reporting and comparison of class areas or class diversity, as completed with the NFI data, will not be required. Investigations are in progress to explore the utility of various spatial support regions for both the labelling protocol and the definition of agreement between the map and reference sources; it is anticipated that some form of a spatial support region will be necessary for both these purposes.

7. Conclusion

The accuracy assessment of large-area land cover products presents many unique challenges to the remote sensing community; however, rigorous validation methods are emerging from the literature as the creation and assessment of large-area land cover products increases. In this paper, we have provided an accuracy assessment framework which details the key decision points that should be addressed in the design of a statistically robust validation of large-area land cover products. These decision points are summarized in table 4. The overall structure of the framework involves the three key components identified by Stehman and Czaplewski (1998): sample design, response design and analysis. Each of these components must be considered in concert as each has impacts on the overall design.

In the context of selecting the appropriate sample design, the target population must be identified, the unit of sampling must be selected, the desired level of accuracy and significance level must be chosen, and the required sample size for hypothesis testing (and the method of testing) must be determined. Most

Table 4. Accuracy assessment framework for large area land cover products.

OBJECTIVE	What is the objective of the accuracy assessment?	• What is the priority: overall map accuracy or individual class accuracy (or both) • Are there rare classes of special interest? • Is precision a priority for specific target classes of interest? • What are the intended applications for the large area land cover product?
SAMPLE DESIGN	Target population	• What is the target population?
	Sample frame	• How will the samples be selected?
	Sample unit	• What to sample?
	Sample size	• What is the target or accepted level of accuracy? • What method of hypothesis testing will be used? • How many samples are appropriate given the answers to the preceding two questions?
	Sample design	• Where to sample (what method of sample selection will be used)?
RESPONSE DESIGN	Evaluation protocol	• What is the spatial support region (SSR) on which the reference land cover evaluation will be determined? • What is the size and shape of the SSR? • How will the SSR be used to determine the reference classification? • What reference data are to be collected?
	Labelling protocol	• How will the reference data be classified for comparison with the image data? • What will be the definition of agreement between the map label and the reference label? • Will the protocol be relevant from the perspective of the end-user?
ANALYSIS AND ESTIMATION	Error matrix	• Report in proportions rather than counts.
	Estimation of accuracy parameters	• What is the appropriate formula to estimate the accuracy parameters based on the sample design?
	Variance	• What is the appropriate formula to estimate the variance for the accuracy parameters based on the sample design?

importantly, the method for selecting the samples must be chosen, considering the requirements of a spatially representative sample, efficiency, and a probability sampling scheme where inclusion probabilities for all samples are known explicitly.

The sample design may be chosen regardless of the actual reference data that will be used to validate the classification. The response design, on the other hand, deals with all the details specifically associated with the reference data. A response design is comprised of an evaluation protocol and a labelling protocol. These protocols determine the size and shape of the SSR, as well as the manner in which the SSR is to be used to determine the reference classification for the sampling unit (e.g. mode, mean). These protocols also detail how the reference data will be classified and what definition of agreement will be used for the validation. Ultimately, the response design identifies what reference data are to be collected. The spatial extent of large-area land cover products typically results in constraints on the types of reference data that can be used.

The final component of the accuracy assessment framework is the analysis of the results and the estimation of the accuracy parameters. An error matrix is a useful mechanism for summarizing results and facilitating the calculation of accuracy measures; it is vital that appropriate methods be used to calculate the accuracy parameters, as the methods vary with the sample design. The reporting of confidence intervals for accuracy measures provides additional information to the user, as does the interpretation of the results in the context of the stated accuracy assessment objective. Large-area land cover products are complex and require validation plans with clearly defined goals and objectives. Furthermore, these plans must identify, document and execute proper statistical procedures. Ultimately, a well-designed and transparent accuracy assessment protocol will ensure that the user applies the large-area land cover product appropriately—with full awareness of the nature and limitations of the product's accuracy.

Acknowledgements

This research is enabled through funding of the Canadian Space Plan of the Canadian Space Agency and Natural Resources Canada (Canadian Forest Service). For details refer to the following URL: http://www.eosd.cfs.nrcan.gc.ca.

References

ANDERSON, J., 1971, Land-use classification schemes used in selected recent geographic applications of remote sensing. *Photogrammetric Engineering and Remote Sensing*, **37**, pp. 379–387.

ANDERSON, J., HARDY, E., ROACH, J. and WITMER, R., 1976, *A land-use/land-cover classification system for use with remote sensor data*, USGS Professional Paper No. 914 Washington, DC.

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

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

ATKINSON, P. and CURRAN, P.J., 1995, Defining an optimal size of support for remote sensing investigations. *IEEE Transactions on Geoscience and Remote Sensing*, **33**, pp. 768–776.

BAUER, M., BURK, T., EK, A., COPPIN, P., LIME, S., WALSH, T., WALTERS, D., BELFORT, W. and HEINZEN, D., 1994, Satellite inventory of Minnesota forest resources. *Photogrammetric Engineering and Remote Sensing*, **60**, pp. 287–298.

CAMPBELL, J.B., 1981, Spatial correlation effects upon accuracy of supervised classification of land cover. *Photogrammetric Engineering and Remote Sensing*, **47**, pp. 355–363.

CARD, D.H., 1982, Using known map category marginal frequencies to improve estimates of thematic map accuracy. *Photogrammetric Engineering and Remote Sensing*, **48**, pp. 431–439.

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

CIHLAR, J. and BEAUBIEN, J., 1998, *Land Cover of Canada Version 1.1.* Special Publication, NBIOME Project (Ottawa, Ontario: Canada Centre for Remote Sensing and the Canadian Forest Service, Natural Resources Canada).

COCHRAN, W.G., 1977, *Sampling Techniques*, 3rd edition (New York: Wiley).

COHEN, W.B., MAIERSPERGER, T.K., SPIES, T.A. and OETTER, D.R., 2001, Modelling forest cover attributes as continuous variables in a regional context with thematic mapper data. *International Journal of Remote Sensing*, **22**, pp. 2279–2310.

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

CONGALTON, R.G., 1994, Accuracy assessment of remotely sensed data: future needs and directions. In *Proceedings, Pecora 12 Symposium*, 24–26 August 1993 (Bethesda, MD: American Society of Photogrammetry and Remote Sensing), pp. 383–388.

CONGALTON, R.G., 2001, Putting the map back in map accuracy. In *Symposium of the American Society for Photogrammetry and Remote Sensing: Remote Sensing and GIS Accuracy Assessment*, 11–13 December 2001 (Bethesda, MD: American Society of Photogrammetry and Remote Sensing).

CONGALTON, R.G. and GREEN, K., 1999, *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices* (Boca Raton, FL: CRC/Lewis Press).

CZAPLEWSKI, R.L., 1994, *Variance Approximations for Assessments of Classification Accuracy*, Research Paper RM-316 (Fort Collins, CO: US Department of Agriculture, Forest Service, Rocky Mountain Forest and Range Experiment Station).

CZAPLEWSKI, R.L., 1998, Accuracy assessments and areal estimates using two-phase stratified random sampling, cluster plots and the multivariate composite estimator. In *Quantifying Spatial Uncertainty in Natural Resources: Theory and Applications for GIS and Remote Sensing*, H.T. Mowrer and R.G. Congalton (Eds) (Chelsea, MI: Ann Arbor Press), pp. 79–100.

CZAPLEWSKI, R.L., 2003, Statistical design and methodological considerations for the accuracy assessment of maps of forest condition. In *Remote Sensing of Forest Environments: Concepts and Case Studies*, M.A. Wulder and S.E. Franklin (Eds) (Boston MA: Kluwer Academic). pp. 115–141.

CZAPLEWSKI, R.L. and PATTERSON, P.L., 2000, Accuracy of remotely sensed classification for stratification of forest and non-forest lands. In *Proceedings of the Second Annual Forest Inventory and Analysis Symposium*, Salt Lake City, UT, 17–18 October, edited by G.A. Reams *et al.*, Forest Service General Technical Report SRS-GTR-47.

CZAPLEWSKI, R.L. and PATTERSON, P.L., 2003, Classification accuracy for stratification with remotely sensed data. *Forest Science*, **49**, pp. 402–408.

EDWARDS JR, T.C., MOISEN, G.G. and CUTLER, D.R., 1998, Assessing map accuracy in a remotely sensed, ecoregion-scale cover map. *Remote Sensing of Environment*, **63**, pp. 73–83.

FOODY, G.M., 2002, Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, **80**, pp. 185–201.

FRANKLIN, S.E., PEDDLE, D., WILSON, B. and BLODGETT, C., 1991, Pixel sampling of remotely sensed digital imagery. *Computers and Geosciences*, **17**, pp. 759–775.

FRANKLIN, S.E. and WULDER, M.A., 2002, Remote sensing methods in large area land cover classification using satellite data. *Progress in Physical Geography*, **26**, pp. 173–205.

FULLER, R., GROOM, G. and JONES, A., 1994, The land cover map of Great Britain: an automated classification of Landsat Thematic Mapper data. *Photogrammetric Engineering and Remote Sensing*, **60**, pp. 553–562.

FULLER, R.M., SMITH, G.M. and DEVEREUX, B.J., 2003, The characterization and measurement of land cover change through remote sensing: problems in operational applications? *International Journal of Applied Earth Observation*, **4**, pp. 243–253.

GILLIS, M.D., 2001, Canada's National Forest Inventory (responding to current information needs). *Environmental Monitoring and Assessment*, **67**, pp. 121–129.

GOODCHILD, M.F., BIGING, G.S., CONGALTON, R.G., LANGLEY, P.G., CHRISMAN, N.R. and DAVIS, F.W., 1994, *Final Report of the Accuracy Assessment Task Force, California Assembly Bill AB1580*, National Center for Geographic Information and Analysis, UCSB.

HANSEN, M.C., DEFRIES, R.S., TOWNSHEND, J.R.G. and SOHLBERG, R., 2000, Global land cover classification at 1 km spatial resolution using a classification tree approach. *International Journal of Remote Sensing*, **21**, pp. 1331–1364.

HOMER, C., RAMSEY, R., EDWARDS JR, T. and FALCONER, A., 1997, Landscape cover-type modeling using a multi-scene thematic mapper mosaic. *Photogrammetric Engineering and Remote Sensing*, **63**, pp. 59–67.

JANSSEN, L. and VAN DER WEL, F., 1994, Accuracy assessment of satellite derived land-cover data: a review. *Photogrammetric Engineering and Remote Sensing*, **60**, pp. 419–426.

JUSTICE, C., BELWARD, A., MORISETTE, J., LEWIS, P., PRIVETTE, J. and BARET, F., 2000, Developments in the 'validation' of satellite sensor products for the study of the land surface. *International Journal of Remote Sensing*, **21**, pp. 3383–3390.

KALKHAN, M.A., REICH, R.M. and STOHLGREN, T.J., 1998, Assessing the accuracy of Landsat Thematic Mapper classification using double sampling. *International Journal of Remote Sensing*, **19**, pp. 2049–2060.

KISH, L., 1967, *Survey Sampling* (New York: Wiley).

KISH, L., 1988, Mulitpurpose sample designs. *Survey Methodology*, **14**, pp. 19–32.

KYRIAKIDIS, P.C. and DUNGAN, J.L., 2001, A geostatistical approach for mapping classification accuracy and evaluating the impact of inaccurate spatial data on ecological model predictions. *Environmental and Ecological Statistics*, **8**, pp. 311–330.

LABA, M., GREGORY, S.K., BRADEN, J., OGURCAK, D., HILL, E., FEGRAUS, E., FIORE, J. and DEGLORIA, S.D., 2002, Conventional and fuzzy accuracy assessment of the New York Gap Analysis Project land cover map. *Remote Sensing of Environment*, **81**, pp. 443–455.

LOVELAND, T.R. and BELWARD, A.S., 1997, The International Geosphere Biosphere Programme Data and Information System global land cover data set (DISCover). *Acta Astronautica*, **41**, pp. 681–689.

LOVELAND, T.R., MERCHANT, J.W., OHLEN, D.O. and BROWN, J.F., 1991, Development of a land-cover characteristics database for the conterminous US. *Photogrammetric Engineering and Remote Sensing*, **57**, pp. 1453–1463.

LOVELAND, T.R., REED, B.C., BROWN, J.F., OHLEN, D.O., ZHU, Z., YANG, L. and MERCHANT, J.W., 2000, Development of a global land cover characteristic database and IGBP DISCover from 1 km AVHRR data. *International Journal of Remote Sensing*, **21**, pp. 1303–1330.

MA, Z., HART, M.M. and REDMOND, R.L., 2001, Mapping vegetation across large geographic areas: integration of remote sensing and GIS to classify multiresource data. *Photogrammetric Engineering and Remote Sensing*, **67**, pp. 295–307.

MAGNUSSEN, S., 2001, Fast pre-survey computation of the mean spatial autocorrelation in large plots composed of a regular array of secondary sampling units. *Mathematical Modelling and Scientific Computing*, **13**, pp. 204–217.

MCGWIRE, K.C. and FISHER, P., 2001, Spatially variable thematic accuracy: beyond the confusion matrix. In *Spatial Uncertainty in Ecology: Implications for Remote Sensing and GIS Applications*, C.T. Husaker, M.F. Goodchild, M.A. Friedl and T.J. Case (Eds) (New York: Springer). pp. 308–329.

MERCHANT, J.W., YANG, L. and YANG, W., 1994, Validation of continental-scale land cover databases developed from AVHRR data. In *Proceedings, Pecora 12 Symposium*, 24–26 August 1993 (Bethesda, MD: American Society of Photogrammetry and Remote Sensing), pp. 63–72.

MOISEN, G.G., EDWARDS, T.C. and CUTLER, D.R., 1994, Spatial sampling to assess classification accuracy of remotely sensed data. In *Environmental Information Management and Analysis: Ecosystem to Global Scales*, J. Brunt, S.S. Stafford and W.K. Michener (Eds) (Philadelphia, PA: Taylor and Francis). pp. 161–178.

MORISETTE, J.T., PRIVETTE, J.L. and JUSTICE, C.O., 2002, A framework for the validation of MODIS land products. *Remote Sensing of Environment*, **83**, pp. 77–96.

MUCHONEY, D., STRAHLER, A., HODGES, J. and LOCASTRO, J., 1999, The IGBP DISCover confidence sites and the system for terrestrial ecosystem parameterization: tools for validating global land-cover data. *Photogrammetric Engineering and Remote Sensing*, **65**, pp. 1061–1067.

NUSSER, S.M. and KLAAS, E.E., 2002, *Final Performance Report to EPA Region 7*, Part II: *Gap Accuracy Assessment Pilot Study* (Ames, IA, USA: Environmental Protection

Agency Contract X997387-01 Final Report, Iowa Cooperative Fish and Wildlife Research Unit, Iowa State University).

NUSSER, S.M. and KLAAS, E.E., 2003, Survey methods for assessing land cover map accuracy. *Environmental and Ecological Statistics*, **10**, pp. 309–331.

REESE, H.M., LILLESAND, T.M., NAGEL, D.E., STEWART, J.S., GOLDMANN, R.A., SIMMONS, T.E., CHIPMAN, J.W. and TESSAR, P.A., 2002, Statewide land cover derived from multiseasonal Landsat TM data: a retrospective of the WISCLAND project. *Remote Sensing of Environment*, **82**, pp. 224–237.

SÄRNDAL, C.E., SWENSSON, B. and WRETMAN, J., 1992, *Model Assisted Survey Sampling* (New York: Springer).

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

SKIDMORE, A.K. and TURNER, B.J., 1992, Map accuracy assessment using line intersect sampling. *Photogrammetric Engineering and Remote Sensing*, **58**, pp. 1453–1457.

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

STEHMAN, S.V., 1996a, Sampling design and analysis issues for thematic map accuracy assessment. In *Proceedings of the 1996 ACSM/ASPRS Annual Convention*, ASPRS Technical Papers, pp. 372–380.

STEHMAN, S.V., 1996b, Cost effective, practical sampling strategies for accuracy assessment of large area thematic maps. In *Spatial Accuracy Assessment in Natural Resources and Environmental Sciences: Second International Symposium*, T.H. Mowrer, R.L. Czaplewski and R.H. Hamre (Eds) (Fort Collins, CO: US Department of Agriculture, Forest Service, Rocky Mountain Forest and Range Experiment Station), pp. 485–492.

STEHMAN, S.V., 1997, Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, **62**, pp. 77–89.

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

STEHMAN, S.V., 2001, Statistical rigour and practical utility in thematic map accuracy. *Photogrammetric Engineering and Remote Sensing*, **67**, pp. 727–734.

STEHMAN, S.V. and CZAPLEWSKI, R.L., 1998, Design and analysis for thematic map accuracy assessment: fundamental principles. *Remote Sensing of Environment*, **64**, pp. 331–344.

STEHMAN, S.V. and CZAPLEWSKI, R.L., 2003, Introduction to special issue on map accuracy. *Environmental and Ecological Statistics*, **10**, pp. 301–308.

STEHMAN, S.V. and OVERTON, W.S., 1996, Spatial sampling. In *Practical Handbook of Spatial Statistics*, S.L. Arlinghaus and D.A. Griffith (Eds) (Boca Raton, FL: CRC Press), pp. 31–63.

STEHMAN, S.V., WICKHAM, J.D., YANG, L. and SMITH, J.H., 2000, Assessing the accuracy of large area land cover maps: experiences from the multi-resolution land-cover characteristics (MRLC) project. In *Proceedings of the Fourth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, G. Heuvelink and M. Lemmens (Eds) (Delft: Delft University Press), pp. 601–608.

STEHMAN, S.V., WICKHAM, J.D., SMITH, J.H. and YANG, L., 2003, Thematic accuracy of the 1992 national land-cover data for the eastern United States: statistical methodology and regional results. *Remote Sensing of Environment*, **86**, pp. 500–516.

STEVENS, D.L. Jr., 1994, Implementation of a national monitoring program. *Journal of Environmental Management*, **42**, pp. 1–29.

STEVENS, D.L. Jr. and OLSEN, A.R., 2004, Spatially-balanced sampling of natural resources. *Journal of the American Statistical Association*, **99**, pp. 262–278.

STONE, T., SCHLESINGER, P., HOUGHTON, T. and WOODWELL, G., 1994, A map of the vegetation of South America based on satellite imagery. *Photogrammetric Engineering and Remote Sensing*, **60**, pp. 541–551.

TØMMERVIK, H., HØGDA, K.A. and SOLHEIM, I., 2003, Monitoring vegetation changes in Pasvik (Norway) and Pechenga in Kola Peninsula (Russia) using multitemporal Landsat MSS/TM data. *Remote Sensing of Environment*, **85**, pp. 370–388.

VERBYLA, D.L. and HAMMOND, T.O., 1995, Conservative bias in classification accuracy assessment due to pixel-by-pixel comparison of classified images with reference grids. *International Journal of Remote Sensing*, **16**, pp. 581–587.

VOGELMANN, J.E., HOWARD, S.M., YANG, L., LARSON, C.R., WYLIE, B.K. and VAN DRIEL, N., 2001, Completion of the 1990s national land cover data set for the conterminous United States from Landsat Thematic mapper data and ancillary data sources. *Photogrammetric Engineering and Remote Sensing*, **67**, pp. 650–662.

WICKHAM, J.D., STEHMAN, S.V., SMITH, J.H. and YANG, L., 2004a, Thematic accuracy of the 1992 National Land-Cover Data for the western United States. *Remote Sensing of Environment*, **91**, pp. 452–468.

WICKHAM, J.D., STEHMAN, S.V., SMITH, J.H., WADE, T.G. and YANG, L., 2004b, *A priori* evaluation of two-stage cluster sampling for accuracy assessment of large-area land-cover maps. *International Journal of Remote Sensing*, **20**, pp. 1235–1252.

WULDER, M. and NELSON, T., 2001, *EOSD Land Cover Classification Legend Report*, version 1. Natural Resources Canada, Canadian Forest Service, Pacific Forestry Centre, Victoria, British Columbia, Canada, March 2001. Available from http://www.pfc.forestry.ca/eosd/cover/EOSD_Legend_Report.pdf

WULDER, M. and SEEMANN, D., 2001, Spatially partitioning Canada with the Landsat Worldwide Reference System. *Canadian Journal of Remote Sensing*, **27**, pp. 225–231.

WULDER, M.A., DECHKA, J.A., GILLIS, M.A., LUTHER, J.E., HALL, R.J., BEAUDOIN, A. and FRANKLIN, S.E., 2003, Operational mapping of the land cover of the forested area of Canada with Landsat data: EOSD land cover program. *Forestry Chronicle*, **79**, pp. 1075–1083.

WULDER, M.A., BOOTS, B., SEEMANN, D. and WHITE, J., 2004a, Map comparison using spatial autocorrelation: an example using AVHRR derived land cover of Canada. *Canadian Journal of Remote Sensing*, **30**, pp. 573–592.

WULDER, M.A., FRANKLIN, S.E. and WHITE, J., 2004b, Sensitivity of hyperclustering and labelling land cover classes to Landsat image acquisition date. *International Journal of Remote Sensing*, **25**, pp. 5337–5344.

WULDER, M.A., KURZ, W.A. and GILLIS, M., 2004c, National level forest monitoring and modeling in Canada. *Progress in Planning*, **61**, pp. 365–381.

YANG, L., STEHMAN, S.V., SMITH, J.H. and WICKHAM, J.D., 2001, Thematic accuracy of MRLC land cover for the eastern United States. *Remote Sensing of Environment*, **76**, pp. 418–422.

ZHU, Z. and EVANS, D.L., 1994, US forest types and predicted percent forest cover from AVHRR data. *Photogrammetric Engineering and Remote Sensing*, **60**, pp. 525–532.

ZHU, Z. and WALLER, E., 2003, Global forest cover mapping for the United Nations Food and Agriculture Organization forest resources assessment. *Forest Science*, **49**, pp. 369–380.

ZHU, Z., YANG, L., STEHMAN, S.V. and CZAPLEWSKI, R.L., 1999, Designing an accuracy assessment for a USGS regional land cover mapping program. In *Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources*, K. Lowell and A. Jaton (Eds) (Chelsea, MI: Ann Arbor Press), pp. 393–398.

ZHU, Z., YANG, L., STEHMAN, S.V. and CZAPLEWSKI, R.L., 2000, Accuracy assessment from the US Geological Survey Regional Land Cover Mapping Program: New York and New Jersey Region. *Photogrammetric Engineering and Remote Sensing*, **66**, pp. 1425–1435.