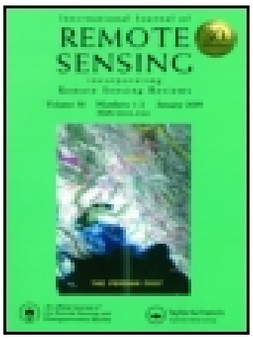

International Journal of Remote Sensing

ISSN: 0143-1161 (Print) 1366-5901 (Online) Journal homepage: www.tandfonline.com/journals/tres20

Sampling designs for accuracy assessment of land cover

Stephen V. Stehman

To cite this article: Stephen V. Stehman (2009) Sampling designs for accuracy assessment of land cover, International Journal of Remote Sensing, 30:20, 5243-5272, DOI: 10.1080/01431160903131000

To link to this article: https://doi.org/10.1080/01431160903131000

Published online: 30 Sep 2009.

International Journal of Remote Sensing
Vol. 30, No. 20, 20 October 2009, 5243–5272

Taylor & Francis
Taylor & Francis Group

Sampling designs for accuracy assessment of land cover

STEPHEN V. STEHMAN*
State University of New York, College of Environmental Science & Forestry, 320 Bray Hall, Syracuse, NY 13210, USA

The accuracy of a land cover classification is the degree to which the map land cover agrees with the reference land cover classification (i.e. ground condition). The basic sampling designs historically implemented for map accuracy assessment have served well for the error matrix based analyses traditionally used. But contemporary applications of land cover maps place greater demands on accuracy assessment, and sampling designs must be constructed to target objectives such as accuracy of land cover composition and landscape pattern. Sampling designs differ in their suitability to achieve different objectives, and trade-offs among desirable sampling design criteria must be recognized and accommodated when selecting a design. An overview is presented of the sampling designs used in accuracy assessment, and the status of these designs is appraised for meeting current needs. Sampling design features that facilitate multiple-objective accuracy assessments are described.

1. Introduction

Land cover data provide critical environmental information for scientific, resource management and policy purposes. Land cover maps are employed in a wide array of applications, including spatial depiction of land cover, calculation of area of the different land cover classes, input into process (mechanistic) or empirical (regression) models, assessment of land cover change, and input in land-use planning decisions such as creation of conservation reserves. 'No land cover classification project would be complete without an accuracy assessment' (Cihlar 2000) succinctly captures the importance of assessing the accuracy of these land cover products. Congalton (1991), Janssen and van der Wel (1994), Stehman and Czaplewski (1998), Congalton and Green (1999), Czaplewski (2003), Foody (2002, 2008), Strahler *et al*. (2006), Wulder *et al*. (2006), and Stehman and Foody (2009) provide general overviews of the basic methodology of accuracy assessment.

An accuracy assessment is based on comparing the map depiction of land cover to the true land cover condition. Typically 'ground truth' is not practically attainable, so accuracy assessments evaluate the map land cover relative to some higher quality determination of land cover. These higher quality data, referred to as 'reference data', are used to produce a 'reference land cover classification' that is compared to the map land cover classification. Since it is too expensive and difficult to obtain the reference land cover classification for the entire region of interest, statistical sampling becomes a critical component of accuracy assessment. A 'sample' is a subset or portion of the region mapped, and a 'sampling design' is a protocol for selecting those locations at which reference data will be collected. The analysis component of an accuracy

*Email: svstehma@syr.edu

International Journal of Remote Sensing
ISSN 0143-1161 print/ISSN 1366-5901 online
http://www.tandf.co.uk/journals
DOI: 10.1080/01431160903131000

assessment then summarizes the data and generates estimates of accuracy. An error or confusion matrix displaying the proportion of area that is correctly classified and misclassified for the different land cover types is a common starting point for analysis. Estimates of overall accuracy and estimates for each individual land cover class are generated from the error matrix.

In this article, the sampling design component of accuracy assessment is singled out for detailed attention. The overarching objective is to provide guidance for choosing a sampling design to assess accuracy of a land cover classification. Planning the sampling design depends on three fundamental features: (1) the accuracy objectives; (2) the desirable sampling design criteria; and (3) the strengths and weaknesses of basic and complex sampling designs relative to these objectives and criteria. The article is structured by these three design planning features. After the basic components of accuracy assessment are introduced (§2), the objectives historically addressed by accuracy assessments are presented. These are the objectives achieved by reporting an error matrix, and summarizing the error matrix information via overall, user's and producer's accuracies (§3.1). Next new accuracy objectives are listed that further assess the utility of a land cover product but require analyses 'beyond the error matrix' (§3.2). These objectives include accuracy of land cover composition and accuracy of landscape pattern. A brief section describing desirable sampling design criteria then follows (§4). An extensive discussion of the strengths and weaknesses of different sampling designs makes up the remainder of the article. This discussion is initiated by describing a framework that organizes basic sampling designs (simple random, systematic, stratified, and cluster sampling) according to decisions focusing on three design features: clusters, strata and selection protocol (§5). Strengths and weaknesses of sampling designs are readily evaluated by focusing on these features. These basic sampling designs are the building blocks of the more complex designs needed to meet an extended set of accuracy objectives. Several of these more complex design options are described and critiqued to appraise the status of sampling design for meeting present-day needs of accuracy information (§6). Sampling design issues related to accuracy assessment of land cover change are reviewed (§7), and the ramifications of the inference framework on sampling are discussed (§8). Problems associated with non-probability sampling are discussed in §9. Developments needed for accuracy assessments to reach their full potential for contemporary use of land cover maps conclude the main text (§10).

2. Components of accuracy assessment

2.1 *Preliminaries*

The starting point for an accuracy assessment is to define a spatial support or assessment unit that will serve as the basis for the location-specific comparison of the map land cover to the reference land cover. The assessment units should form a partition of the region of interest (i.e. the collection of all assessment units should cover the region mapped and the units should not overlap). A pixel is a common choice of assessment unit, but other options include a polygon (e.g. patch of homogeneous land cover), or a regular areal unit such as a 1 km by 1 km block or a 5 by 5 pixel block. The assessment unit need not be equivalent to the minimum mapping unit of the land cover product.

The fundamental sampling design concepts and methods are most easily discussed in the framework of a pixel-based assessment. Difficulties inherent in a pixel-based

assessment are well recognized. For example, although a pixel may be assigned a single map land cover label, the area covered by the pixel may in reality contain more than one land cover class. These 'mixed pixels' create problems when assigning the reference classification to the pixel and defining agreement between the map and reference classifications. Pixel-based assessments are also sometimes criticized because of the potential confounding of location error with classification error when the sample location is not properly matched between the map and the reference data. But these problems do not disappear if a polygon or areal assessment unit is chosen instead. If the assessment unit is something other than a pixel, the analysis and interpretation of the results may differ, but the sampling design concepts and methods are mostly the same. All accuracy assessments, regardless of choice of assessment unit, share many of the same objectives and desirable sampling design criteria. The sampling design considerations are much more strongly determined by the objectives and design criteria than they are by the choice of assessment unit.

The land cover information for each pixel may be the result of a crisp classification in which each pixel is labelled as exactly one class, or a fuzzy classification in which each pixel has a membership value for each class. In this article, it will usually be assumed that both the map and reference classifications are crisp. The basic sampling concepts and methods remain the same for crisp and fuzzy classifications, although specific details of implementation may be different. For example, different rules would be needed to assign pixels to land cover strata for a fuzzy classification from rules applicable to a crisp classification (Stehman *et al*. 2007).

Lastly, it is assumed that the reference data used for accuracy assessment will not be incorporated in the classification process used to create the land cover classification. That is, the accuracy assessment data used for validating the land cover product are independent of the data used to train or develop the classification. If reference data must be used both to create and evaluate the map, cross validation becomes a necessary consideration (see Steele *et al*. (2003) for an overview). The protocol envisioned for this article is that accuracy assessment will be conducted independently of the process used to create the land cover map.

2.2 *Analysis*

In the majority of accuracy assessment applications, both the map and reference classifications are crisp and the descriptive accuracy analyses employ an error matrix to organize the data. Parameters such as overall, user's and producer's accuracies are used to summarize the error matrix information. An example population error matrix is presented in table 1. Each cell of the error matrix represents the proportion of area that falls in the intersection of the map land cover class and the reference land cover class of that cell (i.e. p_{ij} is the proportion of area that is mapped as class i and has reference class j). The cells on the diagonal therefore represent correct classifications, and the off-diagonal cells represent misclassifications. When the population error matrix is constructed in terms of proportion of area, overall accuracy is simply the sum of the diagonal cell entries, user's accuracy is the diagonal entry of each row divided by its respective row marginal proportion, and producer's accuracy is the diagonal entry divided by its respective column marginal proportion. User's and producer's accuracies provide the critically important information on class-specific accuracy, and these accuracy parameters are the complements of commission and omission error probabilities, respectively.

Table 1. Population error matrix and summary parameters for describing accuracy. The cell entry p_{ij} is the proportion of area with map class i and reference class j.

	Reference class					
Map class	1	2	...	k	Row total	User's accuracy
1	p_{11}	p_{12}	...	p_{1k}	$p_{1\cdot}$	$p_{11}/p_{1\cdot}$
2	p_{21}	p_{22}	...	p_{2k}	$p_{2\cdot}$	$p_{22}/p_{2\cdot}$
...	...	...	...	...	...	...
k	p_{k1}	p_{k2}	...	p_{kk}	$p_{k\cdot}$	$p_{kk}/p_{k\cdot}$
Col. Total	$p_{\cdot 1}$	$p_{\cdot 2}$	...	$p_{\cdot k}$		
Producer's accuracy	$p_{11}/p_{\cdot 1}$	$p_{22}/p_{\cdot 2}$	...	$p_{kk}/p_{\cdot k}$		

There is no shortage of other accuracy measures, foremost of which are the chance-corrected measures such as kappa, and map users and producers need to be fully aware of the utility and limitations of these measures (Stehman 1997b, Liu *et al.* 2007). It is also typically the case, at least for large-area land cover maps, that accuracy is reported by region or spatial domain. For example, accuracy might be reported by state (Blackard *et al.* 2008), province (Wu *et al.* 2008), or administrative region (Stehman *et al.* 2003). The error matrix approach is also readily adapted to summarize and describe accuracy of land cover change (Biging *et al.* 1998, Macleod and Congalton 1998, Van Oort 2007).

Gopal and Woodcock (1994) introduced ground-breaking methods for quantifying accuracy when the reference classification is fuzzy and the map classification is crisp. When both the map and reference classifications are fuzzy, Binaghi *et al.* (1999), Lewis and Brown (2001), Pontius and Cheuk (2006), and Pontius *et al.* (2008) extend the error matrix concept to provide descriptive accuracy measures. Entropy-based measures have also been proposed to evaluate agreement between a fuzzy classified map and fuzzy classified reference data (Foody 1995).

Two features common to all of these analyses is that they provide both a measure of overall accuracy as well as measures of class-specific accuracy. The sampling design must therefore provide data that address these features of analysis. In the overall scheme of an accuracy assessment, the objectives determine the analysis, and the analysis strongly influences the choice of sampling design. The main implication of analysis on sampling design is the importance of providing precise estimates of class-specific accuracy. This is the underlying rationale for stratified sampling, one of the basic designs of accuracy assessment.

3. Objectives

3.1 *Basic objectives*

The analyses outlined in the previous section address the basic descriptive objectives typically of interest in accuracy assessment. To re-iterate, these basic objectives are to describe overall accuracy and accuracy of the individual land cover classes. The error matrix and associated summary accuracy parameters are well suited to meet these basic descriptive objectives. Although in concept these basic objectives and error matrix analyses are simple and straightforward, the practical implementation and interpretation of any accuracy assessment has many nuances and caveats that must be considered (Foody 2002, 2008).

3.2 *Extended objectives*

The complexity of information derived from land cover maps and the diversity of applications of these maps dictates extending the objectives of accuracy assessment beyond what can be addressed by a conventional error matrix assessment. These extended objectives include assessing the accuracy of the proportion of area occupied by the different land cover types present within a spatial unit (i.e. land cover composition) and landscape pattern, and accommodating reference data error in the analysis. Each of these objectives will be briefly discussed.

A common application of land cover data is to generate land cover composition for each unit (e.g. 5 km by 5 km block) in a spatial partition of the region of interest (Hollister *et al*. 2004), where land cover composition is the area or proportion of area each land cover class occupies within each of the spatial support units making up this partition. It may be of interest to evaluate accuracy of land cover composition for more than one size of spatial unit (i.e. support). The methodology to assess accuracy of land cover composition is analogous to that of a per-pixel assessment. For a sample of support units, reference data indicating the area or proportion of area occupied by each land cover class are obtained for the sampled units. These reference land cover areas or proportions are then compared to the map areas or proportions (table 2). Mean deviation, mean absolute deviation, root mean square error and correlation are relevant descriptors of accuracy given the quantitative character of land cover composition data. Pontius and Cheuk (2006) and Pontius *et al*. (2008) suggest alternative parameters to quantify accuracy, and Ji and Gallo (2006) provide a general review of parameters potentially applicable in this setting.

Dungan (2006) distinguishes between pixel-based versus feature-based map comparisons. Pixel-based comparison measures include those derived from an error matrix. Features, in contrast, are defined as 'contiguous pixels of constant attribute' (Dungan 2006), as for example a land cover patch or polygon of homogeneous land cover. Feature-based accuracy may be characterized by considering topological relations among features (Hargrove *et al*. 2006), or defined in terms of area and shape of the features, for example, mean patch size, mean perimeter-to-area ratio, or mean fractal dimension of the patches. Clearly an analysis of 'feature accuracy' expands upon the per-pixel assessment focusing only on whether each pixel is correctly classified. Zhan *et al*. (2005) address similar data quality issues only using the term 'objects' instead of 'features'.

Both feature-based accuracy and land cover composition accuracy may be viewed as special cases within a more general assessment addressing accuracy of landscape pattern. Landscape pattern metrics have been constructed to characterize edges of

Table 2. Example data for accuracy assessment of land cover composition. Hypothetical values for per cent area are shown for five primary sampling units (PSUs) or blocks for the class 'Forest'.

Block	Map (%)	Ref (%)	Difference (%)
1	22	26	–4
2	0	5	–5
3	15	10	5
4	55	45	10
5	30	26	4

land cover polygons, patch density, land cover diversity, and contagion or connectedness among patches (McGarigal and Marks 1995). Several software packages have been written to derive landscape pattern metrics from land cover maps (Turner 1990, Baker and Cai 1992, McGarigal and Marks 1995, Riitters *et al*. 1995), which complement metrics that can be extracted from classical GIS overlay functions, for example crops on steep slopes and riparian forest (Wickham *et al*. 1999). Landscape pattern derived from land cover maps is of interest because it provides information that is not directly discernable from the land cover maps alone (e.g. Browder *et al*. 1989, Weber *et al*. 2006). Shao and Wu (2008) emphasize the importance of understanding accuracy of landscape pattern: 'Because most studies of landscape pattern analysis use classified thematic maps based on remote sensing data, the accuracy or uncertainty associated with the maps is critical for our ability to reliably characterize spatial pattern, detect changes, and relate pattern to process'.

Few accuracy assessments of landscape pattern derived from land cover maps have been conducted. Hess and Bay (1997, p. 309) asserted that 'present methods for accuracy assessment are not sufficient for quantifying the uncertainty in landscape indexes that are sensitive to the size, shape, and spatial arrangement of patches'. Hess and Bay (1997) developed a bootstrap analysis incorporating error matrix information to compute confidence intervals for a limited class of landscape pattern indexes (i.e. those that depend only on land cover composition such as landscape diversity indexes). Wickham *et al*. (1997) addressed the uncertainty of landscape patterns derived from land cover maps by introducing classification error, based on the error matrix information, into an existing land cover map and examining the standard deviations of various map measurements over repeated replications of this error generation process. Brown *et al*. (2000) conducted an accuracy assessment of landscape metrics of forest fragmentation, but rather than using a sample of reference data, their assessment was based on computing the fragmentation measures from two overlapping satellite scenes. Shao and Wu (2008) review important results suggesting how classification error may impact accuracy and uncertainty of landscape pattern, but they do not discuss how to directly assess accuracy of landscape pattern.

Accuracy assessment of landscape pattern may be viewed as a map comparison problem where the reference classification is viewed as one 'map' and the land cover map is the other (cf. Dungan 2006). However, most of the map comparison analyses are based on the availability of complete coverage information for both maps and therefore assume a map or complete coverage of the reference classification exists. In the accuracy assessment setting, a complete coverage reference map typically does not exist and only a sample of the reference classification is available. Therefore, the question of how to conduct some of these map comparison analyses from only sample data remains to be addressed.

It is well-recognized that the reference land cover classification is also subject to error (Congalton and Green 1993; Foody 2002, 2008; Powell *et al*. 2004). These errors may be attributable to the interpreter assigning the wrong reference land cover class, a mismatch between the time the reference and map data are obtained, or misregistration between the reference location and corresponding map location. Some of the analyses constructed to accommodate reference data error or to gauge the potential effect of reference data error do not require any special sampling design considerations. For example, Stehman et al. (2003) and Wickham *et al*. (2004a) employed analyses incorporating the map land cover information in a 3×3 pixel window surrounding each sampled pixel. These analyses included a definition of agreement

based on the modal land cover class found in the 3 × 3 pixel window, and reporting accuracy results for the subset of the sample in which the sample pixel was contained within a 3 × 3 window where the map land cover class was the same (i.e. the subset of homogeneous areas of map land cover). Similarly, an analysis that takes into account the possibility that the land cover has changed between the date the map was produced and the date the reference data were obtained would not likely impact the sampling design considerations because the sample focuses on the spatial domain.

The primary source of reference data error that would affect sampling design considerations is the problem of the reference and map locations not being spatially co-registered. To accommodate location error in the analysis, it may be necessary to obtain reference data for a contiguous block of pixels thus allowing analyses for evaluating the impact of a shift of, say, one pixel in various directions. Hagen (2003) provides a rigorous framework for these types of analyses. The approach requires obtaining reference data for blocks of contiguous pixels (e.g. 5 × 5), and the comparison of the reference data to the map classification is conducted within this block and includes a distance weighting function.

In summary, the extended objectives described in this section combined with the basic objectives serve as focal points for planning the sampling design. The overarching objective is to describe accuracy of an end-product land cover map provided to users. This description of accuracy will encompass a broad array of characteristics of the map, beginning with the conventional per-pixel accuracy summarized by an error matrix analysis, but also including accuracy of land cover composition and landscape pattern, and accommodating reference data error in the analysis producing the descriptive results. Still other accuracy objectives may be of interest, but these will not be considered in the sampling design planning deliberations. The objectives not addressed include evaluating the quality of the map at intermediate steps in the classification process, assessing the 'uncertainty' of the potential maps that could have resulted from repeating the classification process (Dungan 2006), characterizing the spatial pattern of classification error, comparing classifiers, and determining and modelling causes of classification error. Once the descriptive accuracy objectives have been satisfied, users of a land cover map will still need to decide whether the map is suitable for their application and determine how error in the map will impact their applications.

4. Desirable sampling design criteria

Choosing a sampling design for accuracy assessment should be guided by the desirable design criteria specified (Stehman 1999). Seven major criteria are proposed.

C1 *The sampling protocol satisfies the requirements of a probability sampling design.* This is the most important design criterion. Probability sampling designs allow the full support of design-based inference (cf. Stehman 2000) to justify the rigour of the accuracy estimates derived from the sample.

C2 *The sampling design must be practical.* A practical design protocol is one in which it is realistic to expect the protocol will be implemented correctly given the conditions and limitations associated with collecting the reference data, and that the analysis of the sample data collected will also be implemented correctly. Correct implementation of the sampling design requires proper randomization at each step of the protocol, and correct implementation of

the analysis requires properly taking into account the inclusion probabilities associated with the sampling design (an inclusion probability is the probability that a specific pixel is selected in the sample, and this probability depends on the sampling design). What is 'practical' is determined to some extent by the abilities of the individuals implementing the design and conducting the analysis.

C3 *The design must be cost effective*. The rationale for this criterion is evident given that funding for accuracy assessment is typically very limited.

C4 *The sample is spatially well distributed* (i.e. 'spatially balanced'). Spatial balance is intuitively appealing because the sample locations are spread throughout the study area without large gaps, and such samples generally tend to result in more precise estimates than non-spatially balanced designs.

C5 *The sampling variability of the accuracy estimates should be small* (i.e. small standard errors). Sampling variability of an estimator is the degree to which the sample-based accuracy estimate would vary over different realizations or random outcomes of the sampling design. Ideally, regardless of the sample selected, the accuracy estimate obtained would be close to the same value, and thus the assessment would be 'repeatable' in the sense that different samples would produce nearly the same accuracy results. This criterion could also be expressed as requiring 'precise' estimates of accuracy, where precision again refers to the repeatability of the accuracy result obtained from the different samples that could potentially arise from the randomized selection process. Generally, increasing the sample size will result in more precise estimates. So in practice, this criterion would compare precision of different sampling designs on an equal cost of sampling basis.

C6 *Sampling variability or precision of the accuracy estimators should be estimated without undue reliance on approximations other than those related to sample size*. Some sampling designs require use of approximate standard errors (i.e. no unbiased estimator of variance is available), as for example, systematic sampling, and other designs such as multi-stage cluster sampling may be so complex that approximating standard errors is a highly desirable practical convenience. Foody's (2008) recommendation to report confidence intervals for accuracy parameters highlights the importance of estimating standard errors to compute these intervals.

C7 *Ability to accommodate a change in sample size at any step in the implementation of the design*. This criterion is valuable because accuracy assessment budgets are often unpredictable and may change after the sampling protocol has been initiated. This criterion can be achieved by judicious choice of a sequential implementation protocol so that if the sampling is terminated prematurely, the part of the sample that has been selected is a legitimate probability sample. Conversely, if it fortuitously turns out that the sample size can be increased beyond what was initially targeted, or if it is necessary to augment the sample size selectively to address a particular objective, the design should readily allow selection of these additional sample elements while still retaining the probability sampling character of the design. Särndal *et al*. (1992, Example 2.2.1, p. 26) describe a sequential selection protocol applicable to simple random and stratified random sampling that allows for terminating the sample at a smaller than planned sample size or adding to the sample beyond the initial target size. This protocol retains the probability sampling feature of the designs.

5. Basic sampling designs

The answers to three questions go a long way toward determining the sampling design for a given accuracy assessment application: (1) Are pixels selected as individual entities or are they grouped into clusters and selected in these clusters? (2) Are the pixels or clusters grouped into strata? (3) Is the sample selection protocol applied to the pixels or clusters (possibly grouped into strata) simple random or systematic? The three questions translate to design choices of whether to use clusters, whether to use strata, and whether to use a simple random or systematic selection protocol. The first two questions address whether and how to group the pixels prior to selecting the sample, whereas the third question directly addresses the randomization protocol for selecting units into the sample. The answers to these three design planning questions depend on the priorities assigned to the objectives and desirable design criteria. Each question is developed more fully in the following subsections.

5.1 *Cluster sampling*

In cluster sampling, spatially contiguous pixels are grouped into clusters (e.g. 5 by 5 pixel blocks) called primary sampling units (PSUs), and the pixels are secondary sampling units (SSUs). In one-stage cluster sampling, all pixels within each sampled PSU would be included in the sample, whereas in two-stage cluster sampling, pixels would be subsampled from the PSUs selected in a first-stage sample of PSUs. Simple versions of two-stage cluster sampling are obtained by selecting a first-stage sample of clusters via either the systematic or simple random selection protocol (§5.3), and then selecting a second-stage sample of pixels within each first-stage cluster via either simple random or systematic selection. The ability to employ different sampling designs at the different stages of a two-stage cluster sampling design affords great flexibility in the implementation of the design. Figure 1 shows a simple example in which the first-stage sample is a systematic sample of clusters followed by a simple random sample of pixels within each first-stage cluster.

The primary rationale for grouping pixels into clusters is to reduce the expense of collecting reference data. For example, if the reference data are obtained by field visits, grouping the sample pixels together reduces travel costs. If the reference data are obtained by interpreting aerial photography, videography, or satellite imagery, grouping the sample pixels yields a cost advantage because it reduces the number of photos or images that would be required. The potential disadvantage of cluster sampling is that if classification error is spatially autocorrelated, cluster sampling may yield larger standard errors relative to an unclustered design of the same cost. Two-stage cluster sampling may considerably reduce costs relative to one-stage sampling while diminishing the variance inflation attributable to positive within-cluster correlation of classification error. Analytic expressions quantifying cost and precision comparisons of one- versus two-stage cluster sampling exist (Cochran 1977), but application of this theory requires quantifying the intra- and inter-cluster correlation of classification error, and reliable estimates of these quantities are often difficult to obtain because relatively large sample sizes are needed.

Historically, the merits of cluster sampling have been considered in terms of the trade-off between precision of the accuracy estimators versus cluster size (Moisen *et al*. 1994; Stehman 1992, 1997a). Revisiting the role of cluster sampling is warranted because the extended accuracy objectives, for example, accuracy of landscape pattern and land cover composition, and net change, require reference data collected for

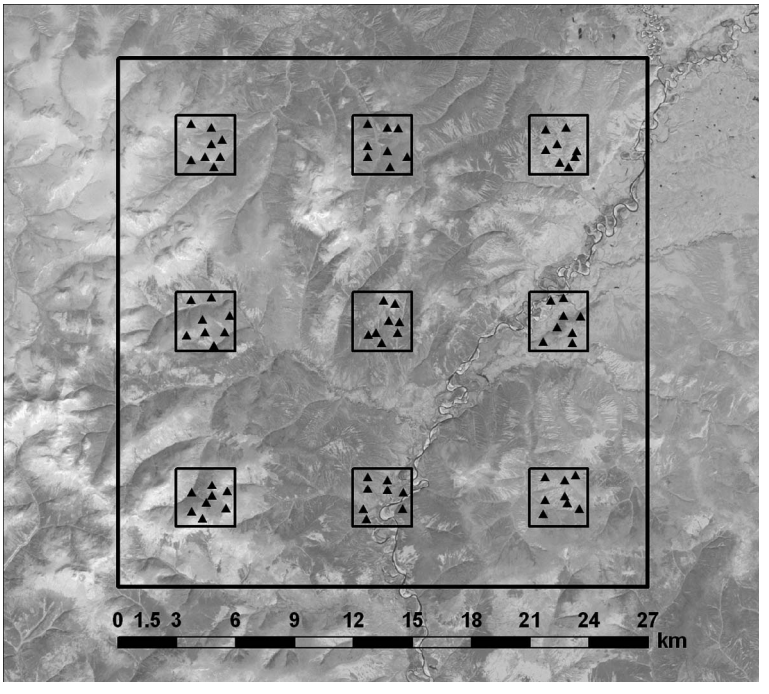

Figure 1. Two-stage cluster sampling design in which the first-stage clusters are selected by a systematic protocol and the second-stage sample of pixels is selected by a simple random sampling protocol within each first-stage sample cluster.

assessment units larger than a pixel (§6). Cluster sampling provides a natural structure for collecting reference data on different size assessment units, a necessity for 'multi-scale' accuracy assessments. The decision of whether to use clusters can no longer be made simply on the basis of precision versus cost, but must now also consider whether the reference data requirements of an extended set of accuracy objectives are better satisfied by cluster sampling.

5.2 *Stratified sampling*

Strata are groups of pixels constructed such that each pixel belongs to exactly one stratum and the strata form a partition of the population of all pixels. Strata are most often constructed based on the map class of each pixel or based on the spatial location of each pixel (i.e. spatial or geographic stratification). Typically the rationale for stratifying by map class is to allocate a disproportionate sample size to rare land cover classes for the objective of precise estimation of class-specific user's accuracy. Geographic stratification can also be used to control sample allocation, as, for example, when the intent is to increase the sample size within one or more relatively small geographic areas. Another rationale for geographic stratification is to construct a spatially balanced sample. By partitioning the region of interest into equal area tessellation cells (e.g. squares or hexagons) and sampling from each cell with equal probability, the sample is assured to be spatially well distributed, much like a systematic sample (de Gruijter *et al*. 2006, p. 77). Geographic stratification can be used to control costs, for example, stratifying by distance from a road and then

sampling more intensively in the near-road stratum (Edwards *et al*. 1998). As sample selection in each stratum is implemented independently of other strata, stratified sampling allows the option to use different sampling designs in different strata. Stratification thus provides flexibility to tailor the design to address different requirements in different strata.

5.3 *Selection protocol*

Once decisions of whether to include clusters and strata have been reached, it still remains to choose the randomization protocol for how the sample units will be selected. Details of implementation of the simple random and systematic selection protocols may be found in basic sampling texts (e.g. Cochran 1977, Särndal *et al*. 1992, Lohr 1999). Simple random sampling (SRS) is a general, all-purpose selection protocol that is readily applied to select a sample of clusters, a sample of units within a cluster (two-stage cluster sampling), or a sample of units within a stratum. Systematic sampling (SYS) is generally motivated because of its ease of implementation in the field and because it achieves the criterion of spatial balance, therefore also tending to produce better precision than SRS. The potential disadvantages of SYS are that it can lead to poor precision if the sampling interval happens to coincide with periodicity in the population (e.g. if classification error is spatially periodic), and it is not possible to construct an unbiased estimator of variance for SYS. Both SRS and SYS share the feature that they are equal probability sampling designs (i.e. each element of the population has an equal probability of being selected in the sample).

Other selection protocols that merit some consideration for accuracy assessment applications are adaptive cluster sampling (ACS) (Thompson and Seber 1996) and general random tessellation stratified (GRTS) sampling (Stevens and Olsen 2004). ACS is a selection protocol constructed to efficiently sample rare, spatially clustered items. Biging *et al*. (1998) noted the potential application of ACS to accuracy assessment of change because change is typically rare and spatially clustered (see §7). GRTS is an innovative, sophisticated selection protocol that achieves spatial balance and retains this property even when considerable non-response is present. A potential application of GRTS would be when field visits are required to obtain the reference classification, but many sample locations are inaccessible because the terrain is difficult or remote or a land owner refuses access. If the reference land cover classification is derived from aerial photography or high resolution satellite imagery, then the non-response problem is usually diminished (frequent cloud cover may be an exception), and it may be simpler to achieve the spatial balance advantage of GRTS by using geographic stratification or SYS. Typically, choosing a selection protocol other than SRS or SYS will incur some increase in complexity of the design or analysis, and this disadvantage should be weighed relative to the advantages that may accrue from the more complex selection protocol.

5.4 *Assembling the design components*

5.4.1 Basic sampling designs. Table 3 lists basic sampling designs that result from the different permutations of choices related to the three design planning questions. Designs *D1*, *D2*, *D3* and *D7* are probably the most familiar of the designs used in accuracy assessment. Design *D4* employs stratification by map land cover class similar to *D3*, but a systematic selection protocol is used to select the sample within each stratum. Designs *D5* and *D6* employ spatial stratification to achieve spatial

Table 3. Basic sampling designs resulting from choices of the three main design components, presence of clusters, presence of strata, and selection protocol (simple random or systematic). Clusters for designs *D9* and *D10* are stratified by dominant land cover type within the cluster.

Designs	Clusters	Strata	Selection protocol (SRS or SYS)
D1: Simple random	No	No	SRS
D2: Systematic	No	No	SYS
D3: Stratified (land cover) random	No	Yes	SRS (within strata)
D4: Stratified (land cover) systematic	No	Yes	SYS (within strata)
D5: Stratified (spatial) random	No	Yes	SRS (within strata)
D6: Stratified (spatial) systematic	No	Yes	SYS (within strata)
D7: Cluster random	Yes	No	SRS
D8: Cluster systematic	Yes	No	SYS
D9: Stratified random cluster	Yes	Yes	SRS (within strata)
D10: Stratified systematic cluster	Yes	Yes	SYS (within strata)

balance, where the spatial strata are equal-area cells partitioning the target region. The typical application of designs *D5* and *D6* is to select a small number of sample pixels from each of a large number of spatial strata. Two versions of *D5* are considered. In the first version (*D5a*), a single pixel is selected from each spatial stratum (this maximizes the spatial balance criterion), and in the second version (*D5b*), the sample size per spatial stratum exceeds 1. Only one case is considered for the spatially stratified design *D6* because a systematic sample of one pixel per spatial stratum would be equivalent to design *D5a*. Defining spatial strata for reporting purposes is not included in table 3. The four cluster sampling designs (*D7* to *D10*) are all one-stage designs, and stratification of clusters (*D9* and *D10*) is assumed to be defined by the dominant map land cover class within the cluster. Design *D10* would employ the same protocol to select the clusters as is used to select pixels in design *D4*. Other designs could be constructed based on slight modifications of the designs listed. For example, within a spatially stratified design, it would be possible to implement simple random selection within some strata and systematic selection within others.

A good sampling design adequately addresses the priority accuracy objectives and satisfies the priority desirable design criteria of a given application. Choosing among the design alternatives requires recognizing the trade-offs of strengths and weaknesses each design has in terms of objectives and desirable criteria. Since all of the sampling designs listed in table 3 are probability sampling designs (*C1*), all permit unbiased or nearly unbiased estimation of the key descriptive parameters overall, user's, and producer's accuracies (Stehman 2001). So bias is not a consideration when deciding which design to implement. Precision is an important, but not the sole criterion used to decide the sampling design as other desirable design criteria may merit equal weight in the design planning deliberations. For example, cost is often a dominant constraint resulting in the imposition of trade-offs among different desirable design criteria.

5.4.2 Criteria-specific comparison of designs. The relative strengths and weaknesses of the basic sampling designs are summarized in table 4. The major generalizations (by design criteria *C2–C7*) reflected in the table 4 ratings are as follows. All but one of the non-cluster sampling designs (*D1–D6*) are rated as strong on the practicality criterion (*C2*). Cluster sampling designs (*D7–D10*) are rated as less practical because the analysis is more complicated by the need to keep track of which cluster each sample pixel was selected from. Design *D4* is rated as neutral on *C2* for reasons explained

Table 4. Relative strengths and weaknesses of basic sampling designs according to desirable design criteria. The criteria are: *C1*) probability sample, *C2*) practical, *C3*) cost, *C4*) spatial balance, *C5*) precise estimates of class-specific accuracy, *C6*) ability to estimate standard errors, and *C7*) flexible to change in sample size. The rating symbols are ●=strength and ○=weakness; absence of a symbol indicates the design is 'neutral' with regard to that criterion. See also section 5.4 in text.

Design	*C1*	*C2*	*C3*	*C4*	*C5*	*C6*	*C7*
D1: Simple random	●	●	○	○	○	●	●
D2: Systematic	●	●	○	●	○	○	○
D3: Stratified (land cover) random	●	●	○	○	●	●	●
D4: Stratified (land cover) systematic	●		○		●	○	○
D5a: Stratified (spatial) random (n_h=1)	●	●	○	●	○	○	
D5b: Stratified (spatial) random (n_h>1)	●	●	○		○	●	●
D6: Stratified (spatial) systematic	●	●	○	●	○	○	○
D7: Cluster random	●		●	○	○	●	
D8: Cluster systematic	●		●		○	○	○
D9: Stratified random cluster	●		●	○			
D10: Stratified systematic cluster	●		●			○	

later. All cluster sampling designs are rated as strong, all other designs weak on the cost criterion (*C3*). This reflects the dominant rationale for using cluster sampling, which is lower cost. Systematic and spatially stratified designs are generally rated as strong on the spatial balance criterion (*C4*), whereas those designs stratified by land cover class are rated as strong on the criterion of precise estimates of class-specific accuracy (*C5*). Any design incorporating a systematic selection protocol is rated as weak on *C6* because an unbiased variance estimator is not available and standard errors will have to be approximated.

In terms of the flexibility criterion (*C7*), designs incorporating a simple random selection protocol are rated as strong and the designs incorporating systematic sampling are rated as weak. If the sample size must be reduced from the target sample size, the systematic structure must be compromised as 'holes' are created by leaving out sample locations. Increasing the sample size requires enhancing or 'densifying' the initial systematic grid. This enhancement can be achieved by decreasing the sampling interval between grid points, for example to half the original distance, but the result is a 4-fold increase in sample size (assuming a square grid). Thus systematic sampling offers very limited flexibility for ease of increasing the sample size by a specified number. Designs incorporating the simple random selection protocol are much more readily reduced or augmented. Designs employing clusters are rated as neutral for *C7* simply because the sample size is less easily controlled by cluster sampling relative to designs without clustering (i.e. entire clusters of pixels must be added or removed rather than individual pixels).

5.4.3 Design-specific strengths and weakness. A few comments on each of the designs helps expand upon the information in table 4. The strengths and weaknesses of the two unclustered, unstratified designs *D1* and *D2* were discussed in §5.3. When class-specific accuracy is a priority objective, the stratified design *D3* becomes a prime candidate. Stratified random sampling using map land cover to form the strata (design *D3*) is one of the most commonly employed sampling designs in accuracy assessment. The many strengths of this design (table 4) help explain its popularity. If

one design had to be designated as the 'universally adequate' accuracy assessment sampling design, design *D3* would likely be the choice.

Implementing a within stratum systematic selection protocol when the strata are the mapped land cover classes (*D4*) is more complex than *D3*. Design *D4* can be implemented by first overlaying a systematic grid over the entire region and then selecting a subset of the pixels from each land cover stratum (Gallego 2005). The initial systematic grid would need to be dense to pick up an adequate number of pixels for the rare land cover classes. As the strict grid pattern of a systematic sample is not preserved once the final stratified sample is selected, this design is rated neutral on the spatial balance criterion (*C4*). The improvement in spatial balance *D4* achieves over *D3* is gained at the expense of ability to estimate standard errors (*C6*) and flexibility to modify the design (*C7*).

The spatially stratified designs (*D5* and *D6*) are constructed to be strong on the spatial balance criterion (*C4*) and in fact these designs would likely only be implemented when *C4* was a very high priority. The spatially stratified designs are weak on cost (*C3*) and precision of class-specific accuracy (*C5*). The trade-offs between the two versions of random selection within spatial strata (*D5*) are that if only one pixel is sampled per stratum (*D5a*), standard errors must be approximated (*C6*) and the flexibility to reduce the sample size (*C7*) is limited because reducing the sample size would require leaving some strata unsampled. If the sample size is more than one pixel per stratum (*D5b*), flexibility to change the sample size becomes a strength because it would be easy to add to or reduce the sample size in each spatial stratum to meet a newly specified overall sample size. Although both *D5* designs are strong on spatial balance, *D5a* would be stronger than *D5b*. Systematic selection within a spatially stratified design (*D6*) has little to offer relative to *D5a* and *D5b* if the sample size per stratum is small. If the spatial strata are reporting regions, there may be more justification for systematic selection within strata.

If cost is a dominant design criterion, the cluster sampling designs *D7* and *D8* are viable options. The choice between the two unstratified cluster sampling designs (*D7* and *D8*) requires considering the usual trade-offs between simple random and systematic selection protocols (§5.3). Whereas other designs employing a systematic protocol are rated as strong on the spatial balance criterion (*C4*), systematic selection of clusters (*D8* and *D10*) receives only a neutral rating because the clusters spatially constrain the sample pixels more than is the case when individual pixels are selected systematically. Since clusters are defined by the spatial proximity of the pixels and not by homogeneity of the land cover within the clusters, a practical difficulty is how to assign clusters to land cover strata when the clusters contain several land cover classes. Several options for assigning each cluster to a stratum exist. For example, the cluster could be assigned to the stratum of the dominant land cover class, or the cluster could be assigned to a stratum based on the presence of a rare class. The latter option would require rules for assigning pixels to strata when multiple rare classes are present. The cluster sampling designs *D9* and *D10*, which are stratified by land cover class, are rated only neutral on the criterion of precise class-specific accuracy (*C5*) because these designs do not allow as much control of the sample size per land cover class as is available for stratified (but not clustered) designs *D3* and *D4*. Designs *D9* and *D10* are rated as neutral for estimating standard errors (*C6*) because the variance estimators will be complicated.

Over the past 30 years, a solid basis of theory and application of sampling design has been established to achieve the traditional error matrix based accuracy

assessments. The basic sampling designs listed in table 3 provide a diverse array of options to choose from depending on the objectives and desirable design criteria, and these sampling designs are the building blocks of more complex designs. The characteristic strengths and weaknesses of each design (table 4) remain present when these designs are incorporated as a component of a more complex design (§6). Constructing more complex sampling designs is motivated by the goal of merging complementary strengths of the basic designs to achieve the extended objectives of accuracy assessment (§3.2). Clusters and strata remain prominent design structures of interest, and the question becomes how to combine these structures to create a design better able to address multiple accuracy objectives and satisfy desirable design criteria.

6. Designs for multi-objective accuracy assessment

6.1 *Cluster sampling*

The reference data requirements of the different accuracy assessment objectives may be organized into three levels of increasing data intensity: pixel (error matrix analyses), larger assessment unit (land cover composition), and larger assessment unit of contiguous pixels (landscape pattern, reference data error). Cluster sampling is well suited to provide a sample of assessment units of different sizes for a multiple-objective accuracy assessment. Cluster sampling allows for concentrating or grouping pixels into a targeted number of sample assessment units (PSUs or blocks), and the nested structure of cluster sampling is an efficient way to collect data useful for analyses requiring different size support units. For example, suppose the objectives stipulate a traditional error matrix assessment at the pixel level (e.g. 30 m pixel), and an assessment of land cover composition accuracy for both a 3 km by 3 km support and a 12 km by 12 km support. A three-stage cluster sampling design using 12 km by 12 km blocks as the primary sampling unit (PSU), 3 km by 3 km blocks as the secondary sampling unit (SSU) nested within the sampled PSUs, and pixels as the final stage sampling unit nested within the sampled SSUs would provide the data required. When the choice of sampling design is driven by the objective of estimating the error matrix and associated accuracy parameters, the primary advantage of cluster sampling is cost. But when the expanded set of accuracy objectives is considered, cluster sampling becomes appealing because it readily accommodates collecting reference data suitable to address these objectives.

Since in one-stage cluster sampling the reference land cover class is obtained for all pixels in the cluster, it is straightforward to compare land cover composition and landscape pattern metrics derived from the reference classification to the corresponding quantities derived from the map classification. Figure 2 illustrates the reference and map land cover for a single sample block. Not only is it easy to visualize any differences between the map and reference classification, but it would be possible to quantify these differences by computing any characteristic of interest such as land cover composition or a landscape pattern metric for both the map and reference classifications of the sample block. One-stage cluster sampling would be consistent with the use of 'maplets' for accuracy assessment (Stoms 1996), with the provision that the maplets (which would be tantamount to clusters) should be selected via a probability sampling protocol. The sample clusters represent a 'block' of pixels or area. As the sample of reference data will not extend beyond the cluster boundary, these cluster boundaries may introduce artificial edge effects affecting the characteristics of land cover patches. For example, when land cover patches are intersected by a

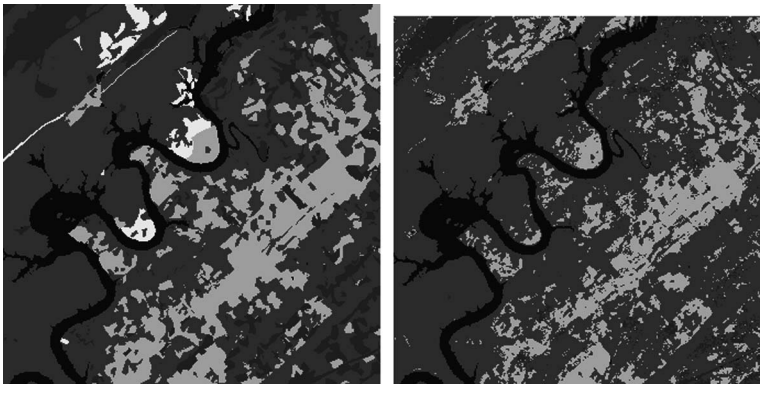

Figure 2. Reference classification (left) and map classification (right) of a single sampled block (primary sampling unit) in one-stage cluster sampling.

cluster boundary, only the portion of the patch inside the cluster will be observed in the reference data protocol, and such truncated patches will affect patch size and patch shape distributions determined from the reference classification. The magnitude of these edge effects will depend on the cluster size and the spatial scale at which landscape patterns are taking place.

Since for one-stage cluster sampling reference data must be collected for all pixels within each sampled cluster, cluster size and number of sample clusters will be strongly affected by cost. For example, if the cluster is a 6 km by 6 km block and the pixel is 30 m by 30 m, a sample of 50 clusters would require the reference land cover class for 2 million pixels. Historically, accuracy assessments have not collected reference data for such large sample sizes. Reducing the cluster size decreases the sampling effort, but the cluster size must still be large enough that it represents a spatial support of practical interest and large enough that edge effects do not adversely impact the analysis.

Two-stage cluster sampling addresses this cost concern because subsampling pixels from within each cluster reduces the sample size of pixels. Two-stage cluster sampling is suitable for estimating composition accuracy because the sample size in each cluster can be specified to be large enough that the composition of each sampled cluster is adequately estimated (although at present there have been no studies specifying a minimally required sample size per cluster). However, the two-stage design shown in figure 1 would not be suitable for estimating accuracy of landscape pattern because the design does not provide reference data for contiguous blocks of pixels within each cluster.

6.2 *Sampling designs with strata and clusters*

Combining the advantages of stratification with cluster sampling is highly desirable, particularly if estimating class-specific accuracy is a high priority objective. The two-stage designs using only the simple random or systematic selection protocols (i.e. without stratification) provide no assurance that rare land cover classes will be well represented in the sample. Consequently, stratification by land cover class must be considered. Combining clusters and strata can be achieved in different ways, and several of these design options are reviewed.

Mayaux *et al.* (2006) implemented a stratified one-stage cluster sampling design in which the clusters were Landsat scenes. Four strata were defined by classifying each cluster according to two attributes, the diversity of all land cover classes within the cluster, and percentage of the cluster covered by selected priority classes. As each land cover class is not defined as a stratum, this stratification allows less control over the sample allocation to specific land cover classes relative to a design that defined each map land cover class as a stratum.

Stehman *et al.* (2003, 2008) and Wickham *et al.* (2004a) implemented two-stage cluster sampling with stratification by land cover class incorporated at the second stage. All pixels within a first-stage sample of clusters (6 km by 6 km blocks) were assigned to the stratum corresponding to the pixel's map land cover class. A second-stage stratified random sample was then selected from all pixels available in the first-stage sampled clusters (e.g. 100 pixels selected per land cover stratum). A PSU may contain sample pixels from several different land cover classes. However, the number of pixels sampled in any given PSU is not fixed. Some PSUs may have only a single sample pixel, whereas others could have 50 or more sample pixels. Wulder *et al.* (2006) proposed a design with a similar two-stage structure, but at the second stage, stratified sampling is implemented within each of the first-stage sample clusters (e.g. 50 pixels per land cover stratum per cluster). This design resolves the problem of small sample size per cluster, but would require a much larger total sample size than the design used by Stehman *et al.* (2003). Miettinen *et al.* (2008) employed a two-stage selection protocol in which 20 satellite scenes were first purposely selected to provide geographic spread followed by a stratified random sample from the pixels present in these 20 satellite scenes. The sample size allocated to each land cover class was proportional to the area each class represented in the aggregate area of all 20 scenes.

Couturier *et al.* (2007) proposed a stratified two-stage sampling design with separate PSU selection protocols for common land cover classes (defined as $> 5\%$ or more of the PSU area) and rare land cover classes. The PSUs (based on aerial photographs) for the common classes are selected via simple random sampling. A separate sample is selected for each rare class. In the rare class protocol, PSUs are sampled with inclusion probability proportional to the areal coverage of the rare class in the PSU, and a second-stage pixel is sampled with probability inversely proportional to the first-stage inclusion probability of the PSU that contains the pixel. The sample of pixels selected for each rare class is constrained to a small number of PSUs so that the total number of PSUs required can be fixed at a pre-specified target (e.g. 54 PSUs in Couturier *et al.*'s example), and sample size for each rare class can also be set at a fixed value, say 100 sample pixels per class.

If the SSU is a pixel, two-stage stratified cluster sampling does not provide the necessary data for assessing accuracy of landscape pattern unless the second-stage sampling design is a cluster sample from within each first-stage sample cluster. For example, if the clusters (PSUs) are 6 km by 6 km blocks, a first-stage sample of these clusters could be followed by a second-stage sample of 0.6 km by 0.6 km secondary sampling units (SSUs) to provide the contiguous blocks of pixels necessary for assessing accuracy of landscape pattern and for constructing analyses of the type proposed by Hagen (2003) to accommodate reference data error in the analysis. To increase the sample size for rare classes, the SSUs could be stratified by map land cover, but stratifying these SSUs by land cover class has the same problems associated with stratifying PSUs in one-stage cluster sampling (§5.4.3). There is also the question of what size SSU to use. Hagen's (2003) analyses accommodating reference data error require a minimum size of a 3 by 3 pixel block, but the analysis is likely improved by

using a 5 by 5 pixel block. An even larger size SSU may be needed to adequately assess landscape pattern accuracy.

A comparison of the relative strengths and weaknesses of these different options combining strata and clusters is provided in table 5. As was the case for comparing the basic designs (table 4), different design options represent trade-offs among the desirable design criteria and ability to satisfy objectives. The practicality criterion (*C2*) favours the one-stage designs (whether stratified or not) and the two-stage unstratified designs because of ease of implementation and analysis relative to the designs incorporating stratification (table 5). The comparison of costs (*C3*) is based on a subjective assessment of the number of sample pixels that would be required to obtain a reasonable accuracy assessment. The one-stage cluster sampling designs (*D11–D13*) will be more costly than two-stage sampling *D14*, *D15*, and *D17* because the one-stage designs require sampling all pixels within each cluster. Two-stage design *D16* selects a stratified sample by land cover class within each PSU, so unless a small number of PSUs are sampled, this approach will be relatively costly. The designs using a block SSU (*D18–D21*) are considered more costly than the designs employing a single pixel as the SSU because more data will be required for these block SSUs. As in table 4, designs stratified by map land cover will be rated higher on the criterion of precise class-specific accuracy (*C5*), and if the stratification is of clusters rather than pixels, the pixel-based stratification is rated as stronger for *C5*.

All of the designs listed in table 5 yield reference data suitable for estimating accuracy of land cover composition. One-stage cluster sampling with the PSUs stratified by land cover class (designs *D12* and *D13*) are rated strong for this objective

Table 5. Strengths and weaknesses of sampling designs combining clusters and strata. Ratings are relative to other sampling designs in Table 5, not relative to the basic designs of Table 4. The design criteria are: *C2*) practical, *C3*) cost, and *C5*) precise estimates of class-specific accuracy, and the extended objectives are *COMP*) accuracy of land-cover composition, *PATT*) accuracy of landscape pattern, *REF*) incorporating reference classification error in the analyses. The rating symbols are ●=strength and ○=weakness; absence of a symbol indicates the design is 'neutral' with regard to that criterion or objective. See Section 6 for descriptions of the sampling designs.

Design	*C2*	*C3*	*C5*	*COMP*	*PATT*	*REF*
One stage						
D11: No strata	●	○	○		●	●
D12: Strata – all land-cover classes	●	○		●	●	●
D13: Strata – Mayaux *et al*. (2006)	●	○		●	●	●
Two stage, SSU = pixel						
D14: SRS both stages	●	●	○		○	○
Two-stage, SSU = pixel stratified by map land cover						
D15: Stehman *et al*. (2003)		●	●		○	○
D16: Wulder *et al*. (2006)		○	●	●	○	○
D17: Couturier *et al*. (2007)	○			●	○	○
Two stage, SSU = 5 by 5 pixel block						
D18: SRS at both stages	●		○			●
D19: SSUs stratified by map land cover				●		●
Two stage, SSU = larger pixel block						
D20: SRS at both stages	●		○		●	●
D21: SSUs stratified by map land cover				●	●	●

(COMP). The unstratified designs (*D11*, *D14*, *D18*, and *D20*) are rated neutral because they will not provide precise composition accuracy estimates for the rare land cover classes. Design *D15* is rated as neutral for the COMP objective because it does not control the number of pixels selected in each first-stage sample cluster (Wickham *et al*. 2004b) and some PSUs therefore may have a sample size of only 1 or 2 pixels. *D17* is rated as neutral because it may result in too few PSUs to provide a precise estimate of the proportion of area for some of the land cover classes.

To obtain a strong rating on the *PATT* and *REF* objectives (table 5), the sampling design must provide data for contiguous blocks of pixels. The one-stage cluster sampling designs (*D11–D13*) achieve this requirement, but two-stage cluster sampling does not unless the SSUs are contiguous pixel blocks (*D18–D21*). The analyses incorporating reference classification error can be applied to small pixel block units (*D18*, *D19*). But if these 5 by 5 pixel SSUs are used to assess landscape pattern, land cover patch truncation effects will likely be more prevalent because of the smaller size of the SSU. For this reason, the designs employing the 5 by 5 pixel SSUs (*D18*, *D19*) are rated as neutral for the landscape pattern accuracy objective.

7. Sampling designs for assessing land cover change accuracy

The ability to detect and quantify change in land cover is one of the key contributions of land cover mapping. Accuracy assessment of change shares many similarities with an assessment of a single date land cover product, but some difficulties are exacerbated in the change assessment and affect sampling design considerations. These include the size of the error matrix for summarizing accuracy of change, the rarity of change, and complications associated with collecting reference data at two points in time. The error matrix and associated accuracy parameters remain prominent features in the analysis of land cover change accuracy. If k land cover classes are mapped, a full error matrix for a change accuracy assessment will have k^2 rows and k^2 columns, including k 'no change' possibilities, one for each class mapped, and $k(k-1)$ change transitions to represent the $(k-1)$ possible land cover transitions for each of the k classes. So even if only seven land cover classes are mapped, the full change accuracy error matrix would be 49 rows by 49 columns.

Since land cover change may affect only a small proportion of the landscape, stratified sampling is an obvious consideration to increase the sample representation of the rare change classes (Biging *et al*. 1998). However, the large number of rows in this error matrix suggests that precisely estimating user's accuracy of all no change and change classes may not be practical because of the cost to obtain adequate sample sizes for each class (some strata may be eliminated on the basis that certain land cover transitions are not possible). The number of strata can be reduced considerably by defining just two strata, change and no change. The obvious disadvantage of this stratification is that within each of the change and no change strata, the sample size for the different classes making up these strata will be proportional to the area each class represents in the classification. Common no change classes and common change classes will be well represented in such a sample, but rare classes will be represented by a small number of samples unless the overall sample size is very large. A compromise choice in terms of number of strata would be to retain all no change classes as strata, and to create a single change stratum for each land cover class (e.g. one stratum for the transition from forest to any other class, one stratum for the transition of wetland to any other class). This would result in k no change strata and k change strata for a total

of $2k$ strata. If for example there are $k = 7$ classes mapped, this would require 14 strata instead of the 49 required by stratifying by all possible change outcomes.

The protocol for collecting reference data in the change assessment strongly affects the options for stratification. If the reference data must be collected in 'real' time, for example, by a ground visit to the sample location, then obviously it will not be possible to stratify by map land cover change. Instead, it will be necessary to stratify by anticipated change as determined by land cover change models or by identifying areas of likely change, for example, urban fringes or forests proximate to roads (Biging *et al.* 1998). If the reference data are available in an archived database (e.g. aerial photography or satellite imagery) that includes the initial date of the change period, then the option to stratify by map land cover change exists because it will be possible to obtain archival reference data from the initial date.

Since change is typically rare but spatially clustered, it is natural to consider using adaptive cluster sampling (ACS), a design specifically constructed for such situations. However, several practical factors hinder use of this design. The advantage of ACS is that it creates the ability to increase the sample size for the rare condition of land cover change by adaptively intensifying the sampling in locations where change is found. If the reference data must be collected in 'real' time, obviously ACS cannot be used because it would be impossible to obtain the reference data for the initial time point at the change locations selected by the adaptive protocol. In other words, because the adaptive feature of the strategy takes place at the end of the change period, the locations selected by the adaptive protocol are not known at the initial time period when the reference data must be collected.

If the reference protocol employs archived data, two practical concerns are control of the final sample size to stay within cost constraints and making sure several types of change, not just the most common changes, are well represented in the sample. The former concern may be addressed by modifications of ACS to terminate the sampling at a specified maximum allowable sample size (Thompson and Seber 1996), but these modifications complicate an already complex design and analysis protocol. Ideally, ACS should also have a feature where it would continue the adaptive protocol for the rarer land cover change types after it had reached a stage where the common change types have already reached a requisite sample size. It is not clear if such a feature of ACS has been developed. A last practical deterrent is that if the change assessment is to be conducted on polygons rather than pixels as suggested by Biging *et al.* (1998), details of how to implement ACS with a polygon unit need to be specified. In the situation where ACS is possible (i.e. archived reference data are available), it would also be feasible to stratify by the map change classes to increase the sample size in targeted change classes. The simplicity and familiarity of a stratified sampling design and analysis are major advantages over ACS in this situation.

The error matrix analysis targets the objective of accuracy of gross change, where gross change is defined as the directional change in land cover (e.g. the proportion of the area changed from the forest class to the developed class) on an individual pixel basis. But accuracy assessment could also target net change, where net change is change in the condition of land cover aggregated over a set of spatial units such as watersheds or 10 km by 10 km blocks forming a partition of the region (Stehman and Wickham 2006). Net change takes into account both gains and losses of a land cover class within the areal unit. For example, an accuracy assessment of net change in a forest would evaluate how well the map represented the change in total area of forest for a population of assessment units. As net change is defined for assessment units larger than a

pixel, the sampling design must provide data that allow for estimating net change (as determined from the reference land cover classification) for a sample of these areal units. The data requirements for a net change assessment are much the same as for land cover composition accuracy. It is not necessary to have contiguous pixel blocks in the sample, but it is necessary to have some areal units that contain enough sample pixels to adequately estimate net change according to the reference classification.

To summarize, sampling design for assessing accuracy of land cover change differs little from sampling design for assessing accuracy of a single date land cover map. Stratifying by the map change classes is almost a necessity to ensure adequate sample sizes for precisely estimating class-specific accuracy. The question of how many of the potentially large number of change classes to define as strata is critical in the design planning. Extending the assessment of change accuracy to include accuracy of net change brings to bear issues of size of assessment unit or support for evaluating net change, and also decisions on how to stratify these support units, a problem noted for composition accuracy assessment. Lastly, if the sampling design is to address simultaneously objectives of accuracy of change as well as accuracy of single date maps that might be constructed for the beginning and end of the change period, what are good ways to choose strata that would be effective for all three assessments (change and the two single date land cover products)?

8. Impact of design-based and model-based inference on sampling design

Design-based inference and model-based inference have been discussed in detail by numerous authors in a variety of fields (Hansen *et al.* 1983, de Gruijter and ter Braak 1990, Särndal *et al.* 1992, Gregoire 1998, Stehman 2000, Valliant *et al.* 2000, de Gruijter *et al.* 2006). In brief, design-based inference is the classical approach to inference in survey sampling. The observation on each sampling unit (e.g. presence or absence of classification error for a pixel, or area of forest mapped for a cluster sampling unit) is regarded as a fixed constant, and the uncertainty of inferring from the sample to the population is attributed to the randomization incorporated in the sampling design (i.e. which units appear in the sample). Estimator properties such as bias and variance are determined by the set of all possible samples that could occur for the chosen sampling design along with the probability of selecting each sample in this set. For example, if the target parameter for estimation is the proportion of pixels in a population of N pixels that possess a certain characteristic (denoted by P) and stratified random sampling is implemented, the variance of the sample-based estimator of P (denoted $\hat{P}$) is defined in terms of how much $\hat{P}$ would vary over all possible stratified random samples that could be chosen. It is not necessary and it would, of course, be impossible actually to select all possible samples, but standard error formulas provide an estimate of the variability of interest from the one sample actually selected. Since estimator properties depend on the sampling design, design-based inference is critically dependent on probability sampling to support inference – the sampling design is everything.

In contrast, model-based inference is predicated on the assumption that the sample is fixed, but an observation taken on each sample unit represents the outcome of a random variable. Estimator properties depend on the model. For example, the variance of an estimator would depend on the model specification of the variance of the random variable observed on each sample unit and the covariance between the random variables of pairs of units. If the analysis is addressing the objective of overall accuracy, for example, a simple model specification might be that each sample

observation is regarded as the outcome of a Bernoulli random variable with probability P that the unit (e.g. pixel) is classified correctly and that the covariance between the variable observed on different units depends on the distance between the units (i.e. a spatial covariance structure where the covariance decreases with increasing distance). The sampling design does not play a critical role in model-based inference although certain design structures such as clusters and strata may be represented in the model. Model-based practitioners may favour balanced samples (Valliant *et al.* 2000) in which sample moments of one or more auxiliary variables match the corresponding moments of the full population, but the sample could be purposely, rather than randomly selected to achieve balance. Randomization in selecting the sample may still be justified to insure impartiality and to avoid unconscious or conscious bias in sample selection (Valliant *et al.* 2000, p. 20), but randomization is not needed to support model-based inference.

De Gruijter *et al.* (2006) provide guidelines for deciding between design- and model-based inferences. Applying these guidelines to accuracy assessment, design-based inference is well-suited to the descriptive accuracy objectives (e.g. estimating overall, user's, and producer's accuracies) and in practice sample sizes are almost always large enough to support design-based inference. A model-based approach becomes a necessity if the objective is to map, predict, or model classification error.

'Optimal sampling design' has different meanings in design-based and model-based inference. In design-based inference, optimization usually targets the objective of obtaining a small standard error for an accuracy estimator of interest. Two examples where optimization may impact practice would be the use of optimal allocation of the sample in stratified random sampling (Cochran 1977, §5.5; Särndal *et al.* 1992, §12.7), and the use of optimal sample size formulas to determine how many primary sampling units and how many secondary sampling units to select in two-stage cluster sampling (Cochran 1977, §10.6; Särndal *et al.* 1992, §12.8). In a model-based inference approach, de Gruijter *et al.* (2006, p. 59) suggest that optimization is designed to find the best sampling pattern, where, for example, the best pattern is defined in terms of estimating a variogram or minimizing the mean squared shortest distance between sample locations (e.g. Tapia *et al.* 2005). The model-based approach allows finding such optimal spatial patterns without requiring randomization in the selection of the sample. However, a probability sampling design may produce a sample close to the desired characteristics, albeit not optimal in the model-based sense. For example, the spatially balanced designs generate samples that resemble the pattern of a design minimizing the mean squared shortest distance between sample locations.

The major ramifications on sampling design of design-based versus model-based inference are: (1) probability sampling is critical to support design-based inference; (2) although a probability sampling design may not be ideal for model-based inference, it does not preclude model-based inference; and (3) a purposive, non-probability sampling design that may yield good model-based inferences will not support design-based inferences. The ability to use probability samples in either a design-based or model-based framework is a persuasive reason to choose probability sampling.

9. Deviations from probability sampling

Sample selection protocols that deviate from probability sampling include non-randomized versions of balanced sampling, non-randomized versions of protocols driven by objectives for model-based analyses, restricting sampling to only

homogeneous regions of land cover, and selecting the sample units because of convenience or ease of access. As mentioned in the previous subsection, randomized protocols exist that would accomplish some of the purposes of balanced sampling (Valliant *et al.* 2000, §§3.4.2, 3.4.3, 3.4.5), and randomized spatially balanced designs can generate samples similar in appearance to samples selected to optimize specifications for a model-based approach.

Restricting the sample to only homogeneous regions of land cover is a more nefarious example of non-probability sampling. A rationale for limiting the sample to homogeneous locations is that this diminishes confounding of classification error and location error (i.e. inability to spatially co-register the map and reference sample location), a phenomenon associated with boundaries between different land cover patches. It is well known that sampling only homogeneous areas will produce overly optimistic accuracy estimates (Hammond and Verbyla 1996), and introducing this bias is rarely an acceptable trade-off to address the location error problem.

Another departure from probability sampling occurs when the sample is restricted to easy-to-access locations. Such a protocol is tantamount to re-defining the population to all conveniently accessed areas in the region mapped. If it is necessary to restrict the sample in this way, a probability sample from the reduced population would certainly be preferred to a non-probability selection protocol and would support design-based inference to the population of conveniently accessible location. The difficult task would be to establish via substantive arguments that these accuracy estimates are representative of the entire area mapped. Stehman (2001) presents additional discussion of these issues. Another form of non-probability sampling is purposive selection of locations judged representative by 'expert' knowledge or because the locations are of special interest. A common example of this practise is selecting locations for collecting data to train the classifier. When such purposive sampling is implemented, design-based inference cannot be invoked to justify the representativeness of the accuracy estimates. Edwards *et al.* (2006) present interesting results indicating that not only does this practise result in a biased estimate of accuracy (usually an overestimate), but it also may result in development of a classifier that generalizes poorly.

On occasion, practical reality will dictate no alternative but to use a non-probability-sampling protocol. For example, if there is simply no possibility of funding for collecting a sample for accuracy assessment, the only option may be to set aside a portion of the existing reference data initially targeted for classifier development and use that set aside sample for accuracy assessment. These training data often will not be selected via a probability sampling protocol. Therefore, even if the holdout sample is selected via a probability sampling protocol, the holdout sample still does not represent a probability sample of the map. Lack of a probability sampling design removes the opportunity to invoke the rigour of design-based inference, and once again establishing representativeness of the estimates for the target region is dependent on subject matter rather than probabilistic arguments.

10. Future directions

A comprehensive accuracy assessment of a land cover classification that includes the traditional error matrix objectives as well as evaluation of accuracy of land cover composition and landscape pattern requires assessment units of several sizes. Historically, accuracy assessments have been designed to provide data for the error matrix objectives, but with few exceptions assessments have not been designed to

provide reference sample data for areal units of the size required to evaluate the extended set of accuracy objectives and analyses. It may be possible to use the data from a pixel-based assessment to evaluate accuracy of these higher-order map characteristics, and this would significantly diminish the data requirements placed on the sampling design. To date, few analyses have been conducted in which pixel-based accuracy assessments have been generalized to larger areal analyses. Hess and Bay (1997), Wickham *et al*. (1997), and Leopold *et al*. (2006) are examples of analyses where this generalization is achieved.

Identifying appropriate support sizes for assessing land cover composition accuracy and net change accuracy is necessary input into sampling design planning deliberations. For example, would the majority of map users require land cover composition accuracy for a 1 km by 1 km support or would a 10 km by 10 km support be more informative? These decisions will differ for different land cover mapping applications, but we need to establish ways to determine what support size or sizes are most relevant for a given land cover map. It would also be helpful to know how land cover composition accuracy and net change accuracy vary with change in support size.

At a more fundamental level, the relative importance of the different accuracy objectives needs to be debated. Given limited resources for accuracy assessment and the fact that a design tailored to achieve one accuracy objective likely diminishes the quality of the assessment of other accuracy objectives (tables 4 and 5), each mapping project will need to determine a relative priority among the traditional error matrix objectives, land cover composition accuracy, and landscape pattern accuracy.

Accuracy assessment can be viewed as a map comparison problem, and many map comparison techniques exist for the situation in which complete coverage maps exist. Applying these same analytic methods when only a sample of reference data is available may not be so simple. For example, quantifying land cover composition is a fairly straightforward analysis for a one-stage cluster sampling design because complete coverage reference data are available for each sampled cluster (Hollister *et al*. 2004). But the analysis of composition accuracy is much more complicated for two-stage cluster sampling because only an estimate of the reference land cover composition is available. As yet, few sample-based methods for estimating landscape pattern accuracy have been developed and neither have sample-based estimators been reported for analyses such as Hagen's (2003) to accommodate reference data error.

The main sampling design questions that must be addressed relate primarily to precision and cost. A good starting point would be to evaluate cost and precision of different options for stratification and different choices of cluster sizes and sample allocation for each of the designs critiqued in table 5. In this 'within design' evaluation, the objective would be to identify the best combinations of features to optimize precision for a fixed cost. The cluster size choice is dictated by the outcome of the decision on the size of areal units that are relevant for accuracy of land cover composition and net change. The stratification questions include whether to stratify at the PSU or pixel level, and how to construct these strata if implemented at the PSU level. These design evaluations would address issues such as for a given cluster size, choice of stratification (or no strata), and allocation of the sample to strata and clusters, what is the precision of the estimators of the traditional error matrix parameters and what is the precision of estimators of composition accuracy for each land cover class? Once this within-design evaluation is completed, the next step would be to compare among the different design options. Sampling theory may be of limited help in these evaluations simply because the variance expressions quantifying precision of

the various accuracy estimators will be very complicated. Empirical comparisons based on simulated sampling from known populations would likely play a major role in the evaluation of different design options.

Another promising direction for research would be to explore the relationship between probability sampling and model-based optimized sampling. The potential to implement probability sampling designs that resemble model-based optimized sampling patterns has been noted with spatially balanced probability samples. Other spatial patterns of samples motivated by model-based optimization results could likely be mimicked by an appropriately chosen probability sampling design. For example, Ritter and Leecaster (2007, figure 1) suggest sampling designs with clusters of sample locations surrounding fixed systematic grid points for the purpose of estimating a semivariogram. It seems likely that such spatial patterns could be reproduced using a probability sampling design that combined systematic and cluster sampling. The ability to construct a sampling design that would serve both design-based and model-based inference well would be a highly desirable development.

11. Discussion and conclusions

The basic sampling designs commonly used for accuracy assessment, simple random, stratified random, cluster, and systematic, are often justified when the primary accuracy objectives are satisfied by the traditional error matrix analyses. These designs also constitute the fundamental structures for the more sophisticated sampling designs needed to meet an expanded set of objectives demanded by the contemporary applications of land cover maps. The major shift in thinking about sampling design is motivated by the desire to go beyond the analyses provided by the error matrix.

The 'perfect' sampling design for accuracy assessment must satisfy multiple objectives (i.e. accuracy of the per-pixel land cover classification, land cover composition, and landscape pattern), provide data useful for analyses accommodating reference data error, and satisfy the priority desirable design criteria, most critically cost-effectiveness. The practical reality is that limited resources will require focusing the design on priority objectives, so the key is to choose an adequate design, not necessarily the perfect design, recognizing the strengths and weaknesses of different designs and understanding the trade-offs among objectives and desirable design criteria. The contrasting data requirements of a traditional per-pixel assessment of accuracy versus a block-oriented assessment of land cover composition or landscape pattern accuracy is a prime example of these trade-offs.

Cluster sampling merits strong consideration when the extended accuracy objectives are included because this design provides data applicable to a broad variety of objectives. One-stage cluster sampling can provide data for all of the proposed accuracy assessment objectives and analyses, but it has serious practical limitations related to cluster size and cost. One-stage cluster sampling also forfeits some control over the sample size per land cover class when stratification is combined with one-stage cluster sampling. Two-stage cluster sampling offers a more cost-effective alternative, but this advantage is gained at the potential expense of reduced ability to assess accuracy of land cover composition and landscape pattern.

Many theoretical and practical details remain to be worked out regarding cost effective, efficient ways to implement these multi-stage, cluster sampling designs. Moreover, there is no substitute for the experience gained by implementing these

designs in practice. Implementation and analysis difficulties often are not revealed until an actual sample is selected and the data analysed. The theory and practice of sampling design for basic accuracy assessment applications has a solid footing in the basic probability sampling designs (table 3), and there exists a wealth of collective practical experience using these basic designs. More complex sampling designs combining features of the basic designs will allow extending accuracy assessments to provide the more comprehensive evaluation of map quality required to address the present uses of land cover maps. Additional theoretical development and practical experience with these more complex designs are needed before we reach the same level of confidence in applying these designs to more demanding accuracy evaluations.

Acknowledgments

The anonymous reviewers are thanked for comments that improved the manuscript. James Wickham reviewed an early draft of the manuscript and provided many helpful suggestions along with additional references. David Selkowitz provided figure 1. This article expands upon a paper presented at the 2008 International Symposium on Spatial Accuracy in the Natural Resources, held in Shanghai. The United States Geological Survey Earth Resources Observation and Science (EROS) Center and SUNY ESF provided travel funds for my conference attendance. I would like to thank Dr Jingxiong Zhang and the conference organizing and scientific committees for a highly productive week of sessions and discussions.

References

Baker, W.L, and Cai, Y., 1992, The r.le programs for multiscale analysis of landscape structure using the GRASS geographical information system. *Landscape Ecology*, **7**, pp. 291–302.

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

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

Blackard, J.A., Finco, M.V., Helmer, E.H., Holden, G.R., Hoppus, M.L., Jacobs, D.M., Lister, A.J., Moisen, G.G., Nelson, M.D., Riemann, R., Ruefenacht, B.D., Salajanu, D., Weyermann, D.L., Winterberger, K.C., Brandeis, T.J., Czaplewski, R.L., McRoberts, R.E., Patterson, P.L. and Tymcio, R.P., 2008, Mapping U.S. forest biomass using nationwide forest inventory data and moderate resolution information. *Remote Sensing of Environment*, **112**, pp. 1658–1677.

Browder, J.A., May, L.N., Rosenthal, A. and Gosselink, J.G., 1989, Modeling future harvest trends in wetland loss and brown shrimp production in Louisiana using thematic mapper imagery. *Remote Sensing of Environment*, **28**, pp. 45–59.

Brown, D.G., Duh, J.-D. and Drzyzga, S.A., 2000, Estimating error in an analysis of forest fragmentation change using North American Landscape Characterization (NALC) data. *Remote Sensing of Environment*, **71**, pp. 106–117.

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

Cochran, W.G. 1977. *Sampling Techniques*, 3rd edn (New York, NY: John Wiley & Sons).

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

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

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

Couturier, S., Mas, J.-F., Vega, A. and Tapia, V., 2007, Accuracy assessment of land cover maps in sub-tropical countries: A sampling design for the Mexican National Forest Inventory. *OnLine Journal of Earth Sciences*, **1**, pp. 127–135.

Czaplewski, R.L., 2003, Accuracy assessment of maps of forest condition: Statistical design and methodological considerations. In M.A. Wulder and S.E. Franklin (Eds) *Remote Sensing of Forest Environments: Concepts and Case Studies*, pp. 115–140 (Boston, MA: Kluuwer Academic Publishers).

De Gruijter, J.J. and ter Braak, C.J.F., 1990, Model-free estimation from spatial samples: A reappraisal of classical sampling theory. *Mathematical Geology*, **22**, pp. 407–415.

De Gruijter, J.J., Brus, D., Bierkens, M. and Knotters, M., 2006, *Sampling for Natural Resource Monitoring* (New York, NY: Springer).

Dungan, J.L., 2006, Focusing on feature-based differences in map comparison. *Journal of Geographical Systems*, **8**, pp. 131–143.

Edwards Jr, T.C., Moisen, G.G. and Cutler, D.R., 1998, Assessing map accuracy in an ecoregion-scale cover-map. *Remote Sensing of Environment*, **63**, pp. 73–83.

Edwards Jr, T.C., Cutler, D.R., Zimmermann, N.E., Geiser, L. and Moisen, G.G., 2006, Effects of sample survey design on the accuracy of classification tree models in species distribution models. *Ecological Modelling*, **199**, pp. 132–141.

Foody, G.M., 1995, Cross-entropy for the evaluation of the accuracy of a fuzzy land cover classification with fuzzy ground data. *ISPRS Journal of Photogrammetry and Remote Sensing*, **50**(5), pp. 2–12.

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

Foody, G.M., 2008, Harshness in image classification accuracy assessment. *International Journal of Remote Sensing*, **29**, pp. 3137–3158.

Gallego, F.J., 2005, Stratified sampling of satellite images with a systematic grid of points. *ISPRS Journal of Photogrammetry & Remote Sensing*, **59**, pp. 369–376.

Gopal, S. and Woodcock, C., 1994, Theory and methods for accuracy assessment of thematic maps using fuzzy sets. *Photogrammetric Engineering and Remote Sensing*, **60**, pp. 181–188.

Gregoire, T.G., 1998, Design-based and model-based inference in survey sampling: appreciating the difference. *Canadian Journal of Forest Research*, **28**, pp. 1429–1447.

Hagen, A., 2003, Fuzzy set approach to assessing similarity of categorical maps. *International Journal of Geographical Information Science*, **17**, pp. 235–249.

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

Hansen, M.H., Madow, W.G. and Tepping, B.J., 1983, An evaluation of model-dependent and probability-sampling inferences in sample surveys, *Journal of the American Statistical Association*, **78**, pp. 776–793.

Hargrove, W.W., Hoffman, F.M. and Hessburg, P.F., 2006, Mapcurves: a quantitative method for comparing categorical maps. *Journal of Geographical Systems*, **8**, pp. 187–208.

Hess, G.R. and Bay, J.M., 1997, Generating confidence intervals for composition-based landscape indexes. *Landscape Ecology*, **12**, pp. 309–320.

Hollister, J.W., Gonzalez, M.L., Paul, J.F., August, P.V. and Copeland, J.L., 2004, Assessing the accuracy of National Land Cover Dataset area estimates at multiple spatial extents. *Photogrammetric Engineering & Remote Sensing*, **70**, pp. 405–414.

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

Ji, L. and Gallo, K., 2006, An agreement coefficient for image comparison. *Photogrammetric Engineering & Remote Sensing*, **72**, pp. 823–833.

Leopold, U., Heuvelink, G.B.M., Tiktak, A., Finke, P.A. and Schoumans, O., 2006, Accounting for change of support in spatial accuracy assessment of modeled soil mineral phosphorous concentration. *Geoderma*, **130**, pp. 368–386.

Lewis, H.G. and Brown, M., 2001, A generalized confusion matrix for assessing area estimates from remotely sensed data. *International Journal of Remote Sensing*, **22**, pp. 3223–3235.

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

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

Macleod, R.D. and Congalton, R.G., 1998, A quantitative comparison of change-detection algorithms for monitoring eelgrass from remotely sensed data. *Photogrammetric Engineering & Remote Sensing*, **64**, pp. 207–216.

Mayaux, P., Eva, H., Gallego, J., Strahler, A.H., Herold, M., Agrawal, S., Naumov, S., de Miranda, E.E., di Bella, C.M., Ordoyne, C., Kopin, Y. and Roy, P.S., 2006, Validation of the Global Land Cover 2000 map. *IEEE Transactions on Geoscience and Remote Sensing*, **44**, pp. 1728–1739.

McGarigal, K. and Marks, B.J., 1995, *FRAGSTATS: Spatial Pattern Analysis Program for Quantifying Landscape Structure*. General Technical Report PNW-GTR-351, USDA Forest Service, Pacific Northwest Research Station, Portland, OR.

Miettinen, J., Wong, C.M. and Liew, S.C., 2008, New 500 m spatial resolution land cover map of the western insular Southeast Asia region. *International Journal of Remote Sensing*, **29**, pp. 6075–6081.

Moisen, G.G., Edwards Jr, T.C. and Cutler, D.R., 1994, Spatial sampling to assess classification accuracy of remotely sensed data. In W.K. Michener, J.W. Brunt, and S.G. Stafford (Eds) *Environmental Information Management and Analysis: Ecosystem to Global Scales*, pp. 159–176 (New York, NY: Taylor and Francis).

Pontius, Jr., R.G. and Cheuk, M.L., 2006, A generalized cross-tabulation matrix to compare soft-classified maps at multiple spatial resolutions. *International Journal of Geographical Information Science*, **20**, pp. 1–30.

Pontius, Jr., R.G., Thontteh, O. and Chen, H., 2008, Components of information for multiple resolution comparison between maps that share a real variable. *Environmental and Ecological Statistics*, **15**, pp. 111–142.

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

Riitters, K.H., O'Neill, R.V., Hunsaker, C.T., Wickham, J.D., Yankee, D.H., Timmins, S.P., Jones, K.B. and Jackson, B.L., 1995, A factor analysis of landscape pattern and structure metrics. *Landscape Ecology*, **10**, pp. 23–39.

Ritter, K.J. and Leecaster, M.K., 2007, Multi-lag cluster designs for estimating the semivariogram for sediments affected by effluent discharges offshore in San Diego. *Environmental and Ecological Statistics*, **14**, pp. 41–53.

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

Shao, G. and Wu, J., 2008, On the accuracy of landscape pattern analysis using remote sensing data. *Landscape Ecology*, **23**, pp. 505–511.

Steele, B.M., Patterson, D.A. and Redmond, R.L., 2003, Toward estimation of map accuracy without a probability test sample. *Environmental and Ecological Statistics*, **10**, pp. 333–356.

Stehman, S.V., 1992, Comparison of systematic and random sampling for estimating the accuracy of maps generated from remotely sensed data. *Photogrammetric Engineering & Remote Sensing*, **58**, pp. 1343–1350.

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

Stehman, S.V., 1997b, 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., 2000, Practical implications of design-based sampling inference for thematic map accuracy assessment. *Remote Sensing of Environment*, **72**, pp. 35–45.

Stehman, S.V., 2001, Statistical rigor and practical utility in thematic map accuracy assessment. *Photogrammetric Engineering & 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 Foody, G.M., 2009, Accuracy Assessment. In T.A. Warner, M.D. Nellis, and G.M. Foody (Eds) *The SAGE Handbook of Remote Sensing*, (London: Sage Publications), pp. 297–309.

Stehman, S.V. and Wickham, J.D., 2006, Assessing accuracy of net change derived from land cover maps. *Photogrammetric Engineering & Remote Sensing*, **72**, pp. 175–185.

Stehman, S.V., Wickham, J.D., Smith, J.H. and Yang, L., 2003, Thematic accuracy of the 1992 National Land-Cover Data (NLCD) for the Eastern United States: Statistical methodology and regional results. *Remote Sensing of Environment*, **86**, pp. 500–516.

Stehman, S.V., Arora, M.K., Kasetkasem, T. and Varshney, P.K., 2007, Estimation of fuzzy error matrix accuracy measures under stratified random sampling. *Photogrammetric Engineering & Remote Sensing*, **73**, pp. 165–173.

Stehman, S.V., Wickham, J.D., Wade, T.G. and Smith, J.H., 2008, Designing a multi-objective, multi-support accuracy assessment of the 2001 National Land Cover Data (NLCD 2001) of the conterminous United States. *Photogrammetric Engineering & Remote Sensing*, **74**, pp. 1561–1571.

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

Stoms, D.M., 1996, Validating large-area land cover databases with maplets. *Geocarto International*, **11**(2), pp. 87–95.

Strahler, A.H., Boschetti, L., Foody, G.M., Friedl, M.A., Hansen, M.C., Herold, M., Mayaux, P., Morisette, J.T., Stehman, S.V. and Woodcock, C.E., 2006, Global land cover validation: Recommendations for evaluation and accuracy assessment of global land cover maps, EUR 22156 EN – DG, Office for Official Publications of the European Communities, Luxembourg.

Tapia, R., Stein, A. and Bijker, W., 2005, Optimization of sampling schemes for vegetation mapping using fuzzy classification. *Remote Sensing of Environment*, **99**, pp. 425–433.

Thompson, S.K. and Seber, G.A.F., 1996, *Adaptive Sampling* (New York, NY: John Wiley & Sons).

Turner, M.G., 1990, Spatial and temporal analysis of landscape pattern. *Landscape Ecology*, **4**, pp. 21–30.

Valliant, R., Dorfman, A.H. and Royall, R.M., 2000, *Finite Population Sampling and Inference: A Prediction Approach* (New York, NY: John Wiley & Sons).

Van Oort, P.A.J., 2007, Interpreting the change detection error matrix. *Remote Sensing of Environment*, **108**, pp. 1–8.

Weber, T., Sloan, A. and Wolf, J., 2006, Maryland's green infrastructure assessment: development of a comprehensive approach to land conservation. *Landscape and Urban Planning*, **77**, pp. 94–110.

Wickham, J.D., O'Neill, R.V., Riitters, K.H., Wade, T.G. and Jones, K.B., 1997, Sensitivity of landscape metrics to land cover misclassification and differences in land cover composition. *Photogrammetric Engineering & Remote Sensing*, **63**, pp. 397–402.

Wickham, J.D., Jones, K.B., Riitters, K.H., O'Neill, R.V., Tankersley, R.D., Smith, E.R., Neale, A.C. and Chaloud, D.C, 1999, An integrated environmental assessment of the U.S. mid-Atlantic region. *Environmental Management*, **24**, pp. 553–560.

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*, **25**, pp. 1235–1252.

Wu, W., Shibasaki, R., Yang, P., Ongaro, L., Zhou, Q. and Tang, H., 2008, Validation and comparison of 1 km global land cover products in China. *International Journal of Remote Sensing*, **29**, pp. 3769–3785.

Wulder, M.A., Franklin, S.E., White, J.C., Linke, J. and Magnussen, S., 2006, An accuracy assessment framework for large-area land cover classification products derived from medium-resolution satellite data. *International Journal of Remote Sensing*, **27**, pp. 663–683.

Zhan, Q., Molenaar, M., Tempfli, K. and Shi, W., 2005, Quality assessment for geo-spatial objects derived from remotely sensed data. *International Journal of Remote Sensing*, **26**, pp. 2953–2974.