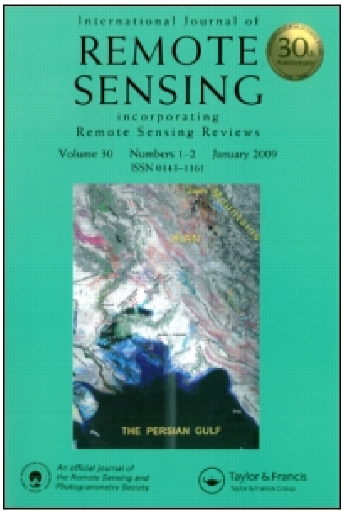

International Journal of Remote Sensing
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713722504

Basic probability sampling designs for thematic map accuracy assessment

Stephen V. Stehman

To cite this Article Stehman, Stephen V.'Basic probability sampling designs for thematic map accuracy assessment', International Journal of Remote Sensing, 20: 12, 2423 — 2441
To link to this Article: DOI: 10.1080/014311699212100
URL: http://dx.doi.org/10.1080/014311699212100

INT. J. REMOTE SENSING, 1999, VOL. 20, NO. 12, 2423–2441

Basic probability sampling designs for thematic map accuracy assessment

STEPHEN V. STEHMAN

SUNY College of Environmental Science and Forestry, 320 Bray Hall, Syracuse, NY 13210, USA; e-mail: svstehma@mailbox.syr.edu

(*Received 9 February 1998; in final form 10 August 1998*)

Abstract. Choosing a sampling design for assessing thematic map accuracy requires the strength of a sampling design to be matched to the objectives and resources available for the accuracy assessment. The criteria to consider when planning the sampling design are that the sample should: (1) satisfy probability sampling protocol; (2) be simple to implement and analyse; (3) result in low variance for the key estimates of the assessment; (4) permit adequate variance estimation; (5) be spatially well distributed; and (6) be cost effective. Several basic probability sampling designs useful for accuracy assessment are reviewed, and recommendations are provided to guide the selection of an appropriate design.

1. Introduction

Accuracy assessments are typically based on a sample of reference locations at which the true or 'reference' land-cover classification is compared to the mapped land-cover classification. The sampling design is the protocol or set of rules for selecting the sampling units that constitute the reference sample. A variety of objectives motivate accuracy assessment. The objectives may include estimating different accuracy parameters, providing information on rare land-cover classes, evaluating contractor compliance, comparing different land-cover classification schemes, evaluating different mapping procedures, and assessing change detection accuracy. No single design is ideal for all objectives. However, some general principles and guidelines can be established to aid in choosing a design that ensures the statistical validity of the assessment and meets practical constraints imposed on the assessment (i.e. cost, data availability and site access). The purpose of this article is to provide an overview of basic probability sampling designs used in accuracy assessment, to review sampling design concepts, and to describe the criteria that should be taken into account when choosing a design.

Within the overall framework of accuracy assessment, the sampling design is one of three basic components, the other two components being the response design and the analysis of the reference data (Stehman and Czaplewski 1998). The response design is the protocol used to obtain the reference land-cover classifications at each sample location. The response design itself may include a statistically valid sampling design, applied within each sampling unit, to assist in assigning a reference class label. For example, if the sampling unit is a polygon, sampling may be employed

International Journal of Remote Sensing
ISSN 0143-1161 print/ISSN 1366-5901 online ©1999 Taylor & Francis Ltd
http://www.tandf.co.uk/JNLS/res.htm
http://www.taylorandfrancis.com/JNLS/res.htm

within each polygon to obtain information relevant to assigning the reference land-cover label to the polygon. Although sampling within the response design component is not discussed further in this article, it is important to recognize the role of the response design in the overall context of accuracy assessment. The response design determines the observation or attribute recorded on each sampling unit, whereas the sampling design determines which units will be selected into the reference sample. Because selecting the sampling unit is independent of the response measured on each unit, the response and sampling designs may be viewed as separate entities. In practice, however, it is advisable to plan the sampling and response designs taking into account simultaneously the options under consideration for each.

The third major component of an accuracy assessment is the data analysis. Analyses often consist of estimating an error matrix and associated summary measures, such as the overall proportion of pixels or polygons correctly classified, user's and producer's accuracies, and probabilities of commission and omission errors. Congalton (1991), Janssen and van der Wel (1994), Stehman (1997b), and Story and Congalton (1986) review some of the commonly used accuracy measures. The sampling design and analysis components are strongly linked because the formulas used to estimate accuracy parameters depend on the sampling design. Accuracy assessment objectives determine the priority of importance of the different accuracy measures, and the sampling design should be chosen to favour precision of those estimates identified as highest priority.

The core sampling design issues can be illustrated by restricting attention to a few fundamental dimensions of the accuracy assessment problem. The reference sample is assumed selected independently of data used to develop the classification, so that the same sample data are not used for both classifier training and accuracy assessment. It is also assumed that the pixels or polygons displayed on the land-cover map are each labelled with one and only one land-cover class, and each reference sample pixel or polygon is also assigned to a single land-cover class. Although relevant to a broad treatment of accuracy assessment, issues such as fuzzy classification approaches (Gopal and Woodcock 1994), confounding of map and location error, determining the reference land-cover classification, and analysing accuracy data are not discussed.

2. Sampling design structures

Several basic structures provide the foundation for the sampling design. An explicit, unambiguous definition of the target population for which map accuracy is desired should be specified first. Usually the target population is the entire region mapped, but the target population may be a subset of the mapped area, for example public lands within the mapped region. The target population may be partitioned into pixels; polygons identified from the classified image, ground visit, or photointerpretation; or some other areal unit such as a 1 km^2 plot. This partitioning determines the basic element of the assessment. In this article, the discussion will be restricted to pixels and mapped land-cover polygons as the basic assessment units. Once the population has been defined and partitioned, a sampling frame must be identified. The sampling protocol is applied to the frame, which is in the form of a list frame of polygons or pixels, or a map frame delineating the boundary of the map region to be assessed. Stehman and Czaplewski (1998) provide further details on the two types of frames and how each type is employed in accuracy assessment.

3. Sampling design criteria

Defining the target population and its partitioning sets the stage for choosing the sampling design. The primary criteria for evaluating a sampling design for accuracy assessment are that the design:

(C1) satisfies probability sampling protocol;
(C2) is simple to implement and analyse;
(C3) has low variance for estimates of the high priority accuracy measures;
(C4) permits adequate variance estimation;
(C5) results in a sample that is spatially well distributed;
(C6) is cost effective.

The probability sampling criterion (C1) is described in detail in §4. The other criteria are discussed in the remainder of this section, and several common sampling designs used in accuracy assessment are reviewed on the basis of these criteria in §5.

The simplicity criterion (C2) applies to both the implementation of the design and the analysis of the resulting data. A simple design is one in which the sampling crews are able to execute the design protocol correctly and to find the reference sampling locations with minimal effort, whether these locations are identified on the ground, on aerial photographs, or on video frames. A simple design is also easily adapted to new objectives and demands placed on the assessment, such as accuracy for a rare land-cover type or accuracy within particular subregions of the map. At the analysis stage, the simplicity criterion pertains to the ease of estimating accuracy parameters and their associated standard errors, estimating accuracy over combined land-cover classes or subregions, and estimating accuracy for a revised land-cover classification scheme. Particularly in large-area accuracy assessments, different users will be interested in different components of accuracy, so the simplicity criterion is critical to multi-user, multi-objective assessments.

The simplicity criterion also applies to the ease with which auxiliary information can be incorporated into the analysis to improve the precision of the estimates. One source of auxiliary information is a land-cover classification based on data other than that used to obtain the true reference classification. For example, land-cover classifications obtained from interpreted aerial photography may be used as auxiliary information for reference classifications derived from ground visits. The ground visits are assumed to provide the true reference classification, but the interpreted aerial photography also provides useful information to improve the precision of the accuracy estimates. The auxiliary information can be incorporated into the analysis via a regression estimator (Stehman 1996a) or a multivariate composite estimator (Czaplewski 1994, 1998). Alternatively, the auxiliary information may be the land-cover area proportions derived from the map itself, with these proportions incorporated into the analysis via a poststratified estimator (Card 1982, Stehman 1996b). Although regression and poststratified estimators can be combined with practically any sampling design, the analysis is much easier if the sampling design is simple. Equal probability designs (see §4) are simple to analyse because all sample units are weighted equally in the analysis. Among the class of equal probability designs, simple random sampling provides the easiest analysis in terms of the variance estimation criterion (C4).

The variance criterion (C3) reflects the desire that the sampling design will result in accuracy estimates with acceptably small standard errors to satisfy project objectives. The need to achieve good precision within the cost constraints of the project strongly influences the design options. Typically, several accuracy parameters will be of interest, and often certain land-cover classes and/or subregions are more important to assessment objectives. Planning should take into account the precision requirements of these different objectives and determine which can be met given the available budget. Resources may not be available to satisfy adequately the precision requirements of all objectives. The variance estimation criterion (C4) pertains to quantifying the uncertainty of the accuracy estimates at the analysis stage. This criterion (C4) encompasses the bias and computational ease of the variance estimator. The variance (C3) of the sampling strategy is the quantity estimated by the sample-based variance estimator (C4).

Because of the spatial nature of land-cover mapping, the spatial distribution of the sample (C5) is a relevant design criterion. Different sampling designs produce different spatial patterns. Intuitively, having a sample geographically dispersed throughout the region is appealing. This intuition is formally substantiated because a spatially well-distributed sample improves the precision of accuracy estimates for geographic subregions by increasing the chance that some sample observations will occur within any subregion of reasonable size. Also, having the sample spread across the region often enhances the precision of estimates for the full target population because it diminishes the effect of the positive spatial correlation identified for classification error (Congalton 1988). If describing or mapping the spatial pattern of classification error is of interest, a spatially well-distributed sample should enhance the quality of the spatial description.

The final criterion is cost (C6). The cost and variance criteria are directly linked: the goal is to obtain precise accuracy estimates at the lowest possible cost. Cost savings are achieved at the design stage by appropriate choice of clusters or strata, or at the estimation stage via variance reduction techniques such as poststratified or regression estimation. The specific values for the cost and variance criteria will be project dependent, so it is not possible to specify generally applicable cost and variance standards. Lower cost is associated with smaller sample size, reduced travel time for field crews, smaller numbers of aerial photographs or other materials necessary for the response design, and any variance reduction feature of the sampling strategy that does not incur an increase in sample size (e.g. poststratified analysis).

These design criteria, not surprisingly, are interrelated. In some cases, satisfying one criterion simultaneously enhances another, as illustrated by the mutual benefits of low variance (C3) and cost-effectiveness (C6), and the simultaneous benefit of good spatial distribution (C5) and low variance (C3). At other times, the design criteria are in conflict. For example, larger sample sizes resulting in lower variance (C3) cost more (C6), and a spatially well-distributed sample (C5) is sometimes more costly (C6) because of increased travel time. As another example, a stratified design motivated by the variance criterion (C3) for class-specific estimation (e.g. user's accuracies) results in more complex estimation formulas (C2). Finally, the advantages of a systematic design being spatially well distributed (C5) and simple to implement and analyse (C2) conflict with the disadvantage of an unbiased variance estimator (C4) not being available. The reason that no single design is universally best is because the different sampling designs were developed to address different criteria. Selecting a sampling design almost always requires compromising among these various criteria, so it is important to recognize the advantages and disadvantages of the different designs.

4. Probability sampling

Because probability sampling (C1) provides the mathematical foundation for statistical inference in sampling, it is perhaps the most important design criterion. A probability sample is defined by two conditions: the inclusion probability for each element of the sample is known, and the inclusion probabilities are non-zero for all elements of the population. An inclusion probability, denoted π_u, is the probability that element u is included in the sample, where u is just an arbitrary indexing label. For example, the inclusion probability for simple random sampling (§5.1) with a sample size of n is $\pi_u = n/N$ for each of the N population elements (either pixels or polygons). Systematic sampling (§5.2) implemented via a square grid with a sampling interval of K in both directions from a randomly located starting pixel results in an inclusion probability of $\pi_u = 1/K^2$ for all N pixels.

Inclusion probabilities are a characteristic derived from the set of all possible samples that could arise from a particular sampling design. Inclusion probabilities are different from selection probabilities, the latter being the draw-by-draw or step-by-step probabilities that arise when implementing the sampling protocol. The distinction between inclusion and selection probabilities is illustrated in figure 1, which shows a systematic sample. When selecting the initial sample pixel of the systematic sample, each of the N pixels has probability $1/N$ of being selected. Once this first sample pixel has been chosen, any pixel a multiple of K pixels away in either the horizontal or vertical direction has a selection probability of 1 (it is certain to be selected in this particular systematic sample), and all other pixels have a selection probability of 0 (no chance of being selected in this particular systematic sample). Selection probabilities depend on a particular realization of the sampling design. Conversely, inclusion probabilities apply to the overall sampling process, which in this case encompasses all possible systematic samples that could result from the different randomly selected starting pixels. Although a particular pixel may not have any chance of being selected in a particular realization of the systematic design, this pixel will have a non-zero probability of being selected in a sample initiated from other starting locations. Consequently, when viewed in the context of all realizations that could result from the design, systematic sampling results in a non-zero inclusion probability ($\pi_u = 1/K^2$) for each pixel. Because inclusion probabilities, not selection probabilities, form the basis for sampling inference, it is necessary to specify the inclusion probabilities to satisfy the probability sampling criterion. Selection probabilities are often easier to determine, but knowing the selection probabilities is not necessarily sufficient to derive the inclusion probabilities.

When π_u is the same for all N population elements, the design is called an *equal probability* sampling design. Simple random, systematic, and one-stage cluster sampling are all equal probability sampling designs (§5). Unequal probability designs have different inclusion probabilities for different elements of the population, as occurs, for example, in stratified random sampling with equal allocation. If n_h is the number of pixels mapped in land-cover class (i.e. stratum) h and a simple random sample of size n_h is selected within that stratum, then the inclusion probability is n_h/N_h for each pixel in stratum h. If n_h is the same for all strata (an equally allocated stratified design), the inclusion probabilities (n_h/N_h) will differ among those strata having N_h different, resulting in an unequal probability design. Unequal inclusion probabilities are valid as long as their presence is recognized by the user and accounted for in the analysis by incorporating proper weighting of each sample unit. This weighted estimation is illustrated by the formulas for stratified sampling presented in Card (1982), Green *et al.* (1993), and Stehman (1995).

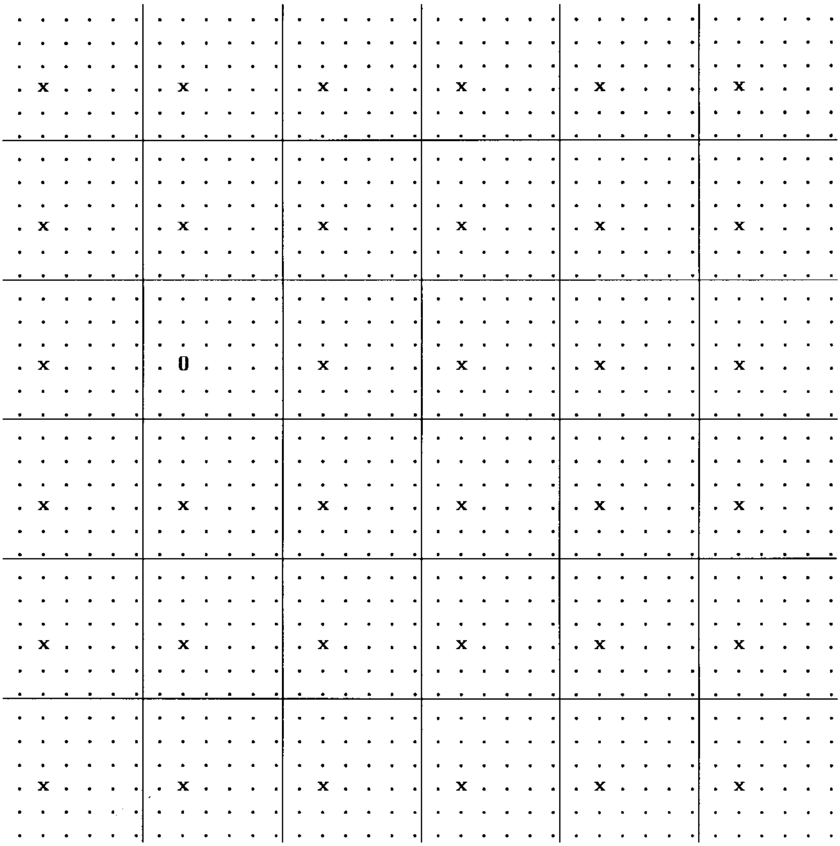

Figure 1. Systematic sampling with sampling interval of $K=6$ in both horizontal and vertical directions. The horizontal and vertical lines are added to illustrate the partitioning of space created by systematic sampling that resembles the partitioning employed in stratified sampling. 0 = randomly selected starting pixel; x = sampled pixel; • = unsampled pixel.

A property of equal probability sampling is that any subgroup of the population is sampled approximately in proportion to its representation in the population. For example, if the sampling unit is a pixel, a land-cover class comprising 25% of the population pixels will comprise about 25% of the sample pixels. A rare land-cover class, say comprising only 0.5% of the population pixels, will be rare in the sample, accounting for approximately 0.5% of the sample pixels. The same feature extends to geographic subregions. If a subregion represents 10% of the total mapped area, an equal probability design applied to pixels will yield roughly 10% of the sample pixels in this subregion. The proportional representation holds for both areas and numbers of pixels. An equal probability sample of polygons from a list frame will also result in proportional representation of polygons for both numbers and areas. That is, if 20% of the map polygons in the population are classified as chaparral, then approximately 20% of the polygons in an equal probability sample will belong to the mapped chaparral class. Chaparral may represent only 5% of the total area of the region, but an equal probability sample yielding proportional representation in terms of numbers translates into proportional representation in terms of area, so about 5% of the sampled area will be classified as chaparral by the map. Any particular equal probability sample selected will deviate from this proportional representation characteristic because of random variation, and only stratified sampling with proportional allocation has the possibility of exact proportional representation for each sample selected. However, the long-run, average behaviour of an equal probability design is proportional representation.

Proportional representation translates to common subgroups of the population being common in the sample. However, we should be careful not to label an equal probability design as 'over-representing' common subgroups and 'under-representing' rarer subgroups. Because of the negative connotations associated with the terms 'under-' and 'over-represented', these labels should be avoided. It too easy to misinterpret 'under-represented' and 'over-represented' as necessarily being a problem, and this may deter users from taking advantage of the analytic simplicity (i.e. unweighted analysis) available from proportional representation.

Non-probability sampling occurs when the sample elements are selected because of their convenience or special interest, or when the sampling protocol is so complex that it is impossible to determine the inclusion probabilities. By definition, if the inclusion probabilities cannot be specified, the protocol is not a probability sampling design. Non-probability sampling results in the loss of the statistical foundation of the assessment, and assumptions must replace the desirable statistical properties assured by probability sampling. For example, it may be assumed that all inclusion probabilities π_us) are greater than 0, or a value of π_u may be assumed for each sampled element when the true value of π_u is unknown. However, it will be nearly impossible to evaluate the validity of these assumptions. To avoid an assessment depending on unverifiable assumptions, probability sampling should be employed.

5. Summary of basic probability sampling designs

The common probability sampling designs can be characterized by whether the selection protocol is simple random or systematic, and by whether one or both of the two primary structures imposed on the population, clusters and strata, is present. Simple random sampling and systematic sampling, applied to a population not partitioned into clusters or strata, represent two of the core sampling designs reviewed (§5.1 and §5.2). If the population is partitioned into clusters, strata, or both, the selection protocol applied to the clusters or to the sampling units within each stratum can be either simple random or systematic. Simple random sampling implemented within each stratum results in stratified random sampling, the third basic design reviewed (§5.3). Finally, designs for sampling a population partitioned into clusters are described in §5.4. These basic designs are then reviewed in §5.5 relative to the criteria listed in §3.

5.1. *Simple random sampling*

Simple random sampling (SRS) may be applied to a list frame of either pixels or polygons, or to an area frame of pixels. The foremost advantage of SRS is simplicity (C2). Estimation and standard error formulas are less complex compared to other designs, and SRS is easy to implement in concept, although finding randomly located sample points in the field is difficult in practice. SRS is readily adaptable to augmenting or reducing the sample size if initial planning considerations over- or underestimate the cost of sampling. To augment a simple random sample of n elements, another simple random sample of size n^* is selected, and the combined sample may be treated as a simple random sample of $n+n^*$ elements. If the augmented sample includes elements already selected in the initial sample, these repeat elements should be discarded. Reducing the sample size is also readily accomplished because a simple random sample of size n' selected from an initial simple random sample of size n is still a simple random sample. As an illustrative example, suppose the population has $N=100$ polygons and the initial sample size is $n=10$. Suppose the polygons are numbered 1–100 in the list frame, and the sampled polygons are numbers 42, 13, 17, 48, 97, 87, 20, 51, 62 and 2. If a decision is made to reduce the sample size to $n=6$, for example because of a decrease in the accuracy assessment budget, a simple random sample of four out of the 10 sample polygons can be deleted to produce a simple random sample with $n=6$. The first six polygons selected in the sequence identified for the simple random sample, polygons 42, 13, 17, 48, 97 and 87 represent a legitimate simple random sample of size $n=6$. As long as the polygons are sampled in the sequence in which they appear in the selection protocol, this sample reduction option remains viable. Conversely, if the polygons are visited in a sequence established by proximity or convenience, the first six polygons sampled in such a sequence no longer represent a simple random sample because this subset of the initial SRS design is not obtained via a SRS protocol.

SRS rates poorly on the variance criterion (C3). For estimating overall accuracy, systematic sampling will generally be more precise than SRS (§5.3), and cluster sampling may result in smaller standard errors than SRS depending on the spatial pattern of classification errors (§5.4). Although SRS incorporates no auxiliary information (e.g. strata or clusters) in the sampling design to enhance precision, it readily accommodates auxiliary information via post-stratified and regression estimators, so cost-effectiveness (C6) may be improved at the estimation stage. These estimation techniques are available, but are more complex for other sampling designs. Because of its equal probability feature, SRS may not produce adequate sample sizes in rare land-cover classes or small subregions to provide estimates with acceptable standard errors. A stratified design, with the rare classes or small subregions defined as strata, would be superior to SRS for estimating accuracy of these subsets of the population. SRS usually produces a sample that is not spatially well distributed (C5), so both systematic sampling and stratified (by geographical region) random sampling are better when a spatially well-distributed sample is a design priority.

5.2. *Systematic sampling*

When the sampling unit is a pixel, a systematic design is implemented by selecting the first (starting) pixel with equal probability assigned to all N pixels. Sampling every Kth pixel in both horizontal and vertical directions from the random starting pixel forms the square sampling grid of pixels (figure 1). A different sampling interval may be chosen for the horizontal and vertical directions resulting in a rectangular rather than a square grid. Systematic sampling is more readily implemented from an area frame than a list frame because it is difficult to construct a list frame that retains the two-dimensional spatial distribution of the pixels or polygons. Implementing a systematic sample in the field is simple because once the random starting position is established, subsequent units are located a fixed distance and direction from that starting position.

Ease of implementation (C2) and good spatial coverage (C5) are reasons systematic sampling has such widespread use in sampling practice. Because it is an equal probability sampling design, systematic sampling shares with SRS all the advantages and disadvantages of equal probability sampling: small subregions and rare land-cover types will be rare in the sample. Analysing the data is relatively simple (C2), and post-stratified or regression estimation may be combined with the systematic design to create an efficient sampling strategy (C3). A common misconception persists that systematic sampling leads to biased estimates of accuracy parameters, but the equal probability sampling feature shared by simple random and systematic designs suggests that systematic sampling estimates are no more susceptible to bias than the corresponding estimates for SRS. Unbiased estimation of variance (C4) is problematic.

Systematic sampling differs from SRS in that it is more difficult to augment or reduce while still retaining the regular grid pattern. For example, to increase the sample size, only certain sample size enhancements will retain the existing systematic grid structure. A systematic sample may be augmented by selecting a simple random sample of additional sampling units, or reduced by selecting a simple random sample from the initial systematic sample, but in both cases, the resulting final sample no longer possesses a pure systematic structure. The analysis of the reduced or augmented systematic sample will be more complex because of the mixture of the two designs employed.

The variance of systematic sampling depends on the spatial pattern of classification errors. In the unlikely event that the classification errors are randomly distributed in space, SRS and systematic sampling have equal variance. As a general guideline, systematic sampling is more precise than SRS when neighbouring pixels or polygons in the population have similar characteristics. That is, if $y_i = 1$ when pixel or polygon i is classified correctly and $y_i = 0$ otherwise, systematic sampling works well if neighbouring elements in the map or neighbouring polygons in a list frame have similar y values. Consequently, if classification errors tend to cluster spatially in the population, systematic sampling will usually have lower variance than SRS because the within-sample variability will be high, a desirable feature for good precision of systematic sampling (Cochran 1977, p. 208). Cochran (§§8.4, 8.6 and 8.9) also compares precision of systematic to stratified sampling. For example, the comparison in Cochran's §8.6 shows that stratified random sampling is superior to systematic sampling in the presence of a linear trend (i.e. an increasing pattern of classification error along a one-dimensional axis of the study region). But from Cochran's figure 8.2 (p. 215), it is clear that the stratified design considered has only one sample observation per stratum. Thus, while stratified sampling may be more precise than systematic sampling when a linear trend in classification error is present, the stratified design that is the basis of the comparison (one sample unit per stratum) suffers from the same inability to estimate variance that is present with systematic sampling. Further, Cochran's results apply to a geographic stratification and not to the commonly used stratification by land-cover class.

Much of the focus in the remote sensing literature has been on when systematic sampling should not be used. The special case in which this concern is potentially justified is when periodicity in the classification errors is likely. If the spatial distribution of errors is periodic, *and* the sampling interval of the systematic design is in phase with this periodicity, then systematic sampling will have high variance relative to SRS. Several unfortunate circumstances must coincide for systematic sampling to perform poorly on the variance criterion (C3). Not only must periodicity exist in the spatial distribution of the errors, but this periodicity must go unnoticed at the planning stage, and the sampling interval must also then coincide with the periodicity.

Situations in which periodicity is likely to be present will often be known. For example, ridge and valley topography, a strong alignment of agricultural fields, or seam lines formed when spatially combining images to create a mosaic may lead to periodicity in the spatial distribution of error. Choosing a sampling interval in phase with these periodic features would consequently be poor planning. Without knowing the spatial distribution of classification error for the entire region, some risk is always present with a systematic sample because of the possibility, however small, that periodicity in the error distribution is indeed present. But what is often overlooked in the discussions of when not to use systematic sampling is that, for the more typical spatially clustered pattern of classification error, systematic sampling may yield a much smaller variance than SRS. To not take advantage of systematic sampling in these cases would be as serious a design flaw as it is to use systematic sampling in those rare cases in which the sampling interval coincides with unknown periodicity. The relationship between the precision of systematic sampling and spatial pattern is illustrated by simulation results in Stehman (1992) and Moisen *et al.* (1994).

Systematic unaligned sampling (Cochran 1977, p. 228) is motivated by the desire to reduce the influence periodicity has on systematic sampling with a regular sampling grid. Systematic unaligned sampling (figure 2) may be viewed as a compromise choice between systematic and simple random sampling. If periodicity is present, systematic unaligned sampling is less susceptible than systematic sampling to the loss of precision resulting when sampling is in phase with population periodicity. Conversely, systematic unaligned sampling diminishes the advantage of systematic sampling when the sampling interval and spatial pattern of errors favours systematic sampling over SRS (Stehman 1992).

Because an unbiased estimator of variance (C4) is not available for systematic sampling, variance estimator formulas appropriate for SRS are used to approximate the systematic sampling variance (Cochran 1977, p. 224). If the spatial distribution of misclassification errors is random so that simple random and systematic sampling have the same variance (C3), then this variance approximation is unbiased. However, because some spatial pattern of the classification errors is almost always present, the variance approximation is usually biased. If systematic sampling has smaller variance than SRS, the SRS variance approximation will overestimate the true systematic sampling variance; conversely, if systematic sampling has higher variance than SRS, the variance approximation will be an underestimate (Särndal *et al.* 1992, §3.4.4) [here it is important to distinguish between the *variance* of the design (C3) and the *estimate* of that variance (C4)]. Knowing the conditions under which the variance of systematic sampling differs from that of SRS allows us to assess which situation probably exists in a practical application. When considering use of systematic sampling, the importance of unbiased versus approximate variance estimation must be determined. Systematic sampling is a viable option if approximate variance estimates are acceptable, and this is typically the case because obtaining accuracy estimates with low variance (C3) via a simple design (C2) is usually more important than unbiased estimation of variance (C4).

Systematic sampling may also be applied to a population partitioned into polygons by selecting those polygons in which a systematic sampling grid point falls. The polygons may be defined by the classified image or interpreted from ground data, aerial photography, or videography. The systematic design can be applied even if all polygons in the population are not delineated, because only those polygons surrounding a systematic grid point need to be identified. The systematic protocol

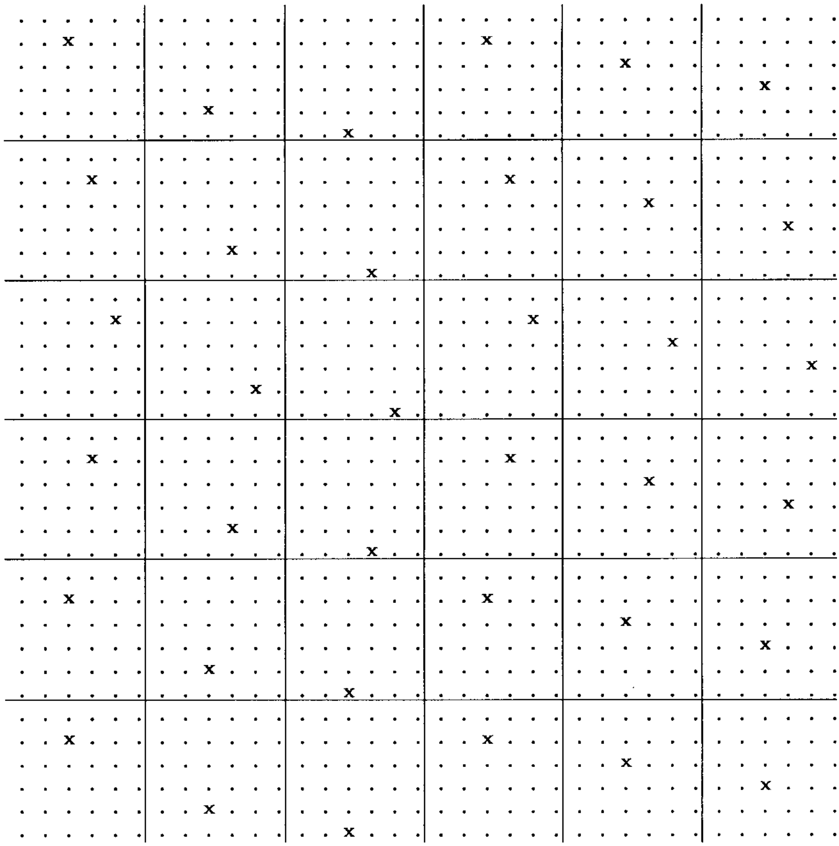

Figure 2. Spatial pattern of systematic unaligned sampling. The horizontal and vertical lines are added to illustrate the partitioning of space created by systematic unaligned sampling that resembles the partitioning employed in stratified sampling. x = sampled pixel; · = unsampled pixel.

applied via this area frame approach achieves a better spatial distribution (C5) than simple random sampling of polygons from a list frame. Larger polygons have higher inclusion probabilities in the area frame protocol. This creates a more complicated analysis because estimates must be weighted properly to account for these unequal inclusion probabilities. The problem of unbiased estimation of variance (C4) is also still present because of the systematic nature of the sample.

5.3. *Stratified sampling*

Stratification is a structure imposed upon the population by assigning each pixel or polygon in the population to one and only one stratum prior to selecting the sample. Sampling then proceeds independently (i.e. separately) among strata, and each stratum may have a different sampling design. Commonly a simple random sample is selected in each stratum, and this design is called stratified random sampling. Stratification is motivated primarily by objectives focusing on stratum-specific estimates of accuracy. The strata are typically the mapped land-cover classes, or geographic subregions. For example, if the objectives require precise estimates of user's accuracy for each mapped land-cover class, then defining the mapped classes as strata and using a stratified design is a good option. Stratified sampling can be used to ensure that the stratum-specific sample sizes are adequate to satisfy the precision requirements (C3) for each stratum. It is not necessary to identify a subgroup (e.g. a geographic subregion or a land-cover class) as a stratum to obtain estimates for this subgroup. Rather, stratification is necessary to control the sample size in these subgroups. Subpopulation or domain estimation (Cochran 1977, §2.12 and §2.13) may be employed to estimate accuracy parameters for subgroups of the population not identified as strata.

When stratified sampling is selected to reduce the standard error for estimating overall accuracy, the gain in precision over SRS is greatest when the proportions differ markedly among strata (Cochran 1977, p. 109). If the strata are mapped land-cover classes, stratified sampling is advantageous over SRS when the land-cover classes show large differences in accuracy. If the strata are different geographic regions, the accuracy in the various regions must differ considerably for stratified sampling to improve precision relative to SRS.

In multi-purpose accuracy assessments, selecting strata to serve the multitude of potential objectives is difficult. Stratifying on more than one attribute may result in a large number of strata if each attribute has many different levels. For example, stratifying by both geographic region and land-cover class may lead to numerous strata if either the number of land-cover classes or geographic regions is large. Unless resources are available for a very large sample, adequate sample sizes will not be available in many of these cross-classified strata to provide precise estimates of within-stratum accuracy parameters (cf. Edwards *et al.* 1998). Objectives may require estimating overall accuracy and stratum-specific accuracy within the same project. These estimates often present a conflict among design criteria. For example, selecting strata and allocating sampling effort for the objective of stratum-specific estimates, usually via equal allocation, results in higher variance for estimates of overall accuracy and producer's accuracy (Stehman 1996b). A compromise solution must be determined by choosing which objective has higher priority.

Many users are less familiar with the estimation formulas needed for stratified designs, and the stratified formulas (cf. Green *et al.* 1993, Stehman 1995) are more complex (C2) than those of simple random or systematic sampling. Estimates combining data across strata, such as overall accuracy, producer's accuracies for a design stratified by mapped land-cover, and regionwide accuracy for a design stratified by geographic subregions, must weight the data because of the unequal inclusion probabilities that result from stratified sampling with any allocation other than proportional allocation. Sometimes strata originally identified are later discarded because the objectives motivating construction of the map changed, or because users do not share a common view of the importance of various subgroups. The original stratified design may then become a liability because having controlled the sample size in subgroups (strata) that are no longer of interest may result in a sample that is poorly distributed among the subgroups that have now become the focus of the assessment. Those new subgroups having only a few sample elements will have imprecise estimates. A common example in which interest in subgroups changes is when land-cover classes are combined within a hierarchical classification scheme. The stratification imposed by the sampling design still affects the analysis, so estimates will require more complicated weighted formulas to combine data from the original strata. Stratification is best reserved for those situations in which the identified strata are nearly certain to retain their importance.

Geographic stratification is motivated by different objectives than stratifying by mapped land-cover. Although the gain in precision from geographic stratification relative to SRS is usually modest (Cochran 1977, p. 102), geographic strata distribute the workload more evenly among administrative units associated with the strata. Stratifying by proximity to a road is a version of geographic stratification that may be employed to reduce travel costs (C6) (cf. Edwards *et al.* 1998). The practical problem of a large number of strata (C2) becomes a concern when other stratification attributes are combined with the proximity-to-road strata.

Geographic stratification may also be employed to improve the spatial distribution (C5) of the sample by employing many small geographic strata. For a target sample size of n, n strata may be formed, and one sampling unit selected at random within each stratum (figure 3). The n spatial strata are sampled independently. The spatial distribution advantage of this stratified design is similar to that of systematic and systematic unaligned sampling, but the stratified design is less susceptible to any potential precision loss attributable to periodicity. Similar to systematic sampling, geographic stratification can reduce the variance of estimates of overall accuracy (C3). If the geographic strata have the same size, this design is an equal probability sampling design, and therefore simple to analyse (C2). If only one sample unit per

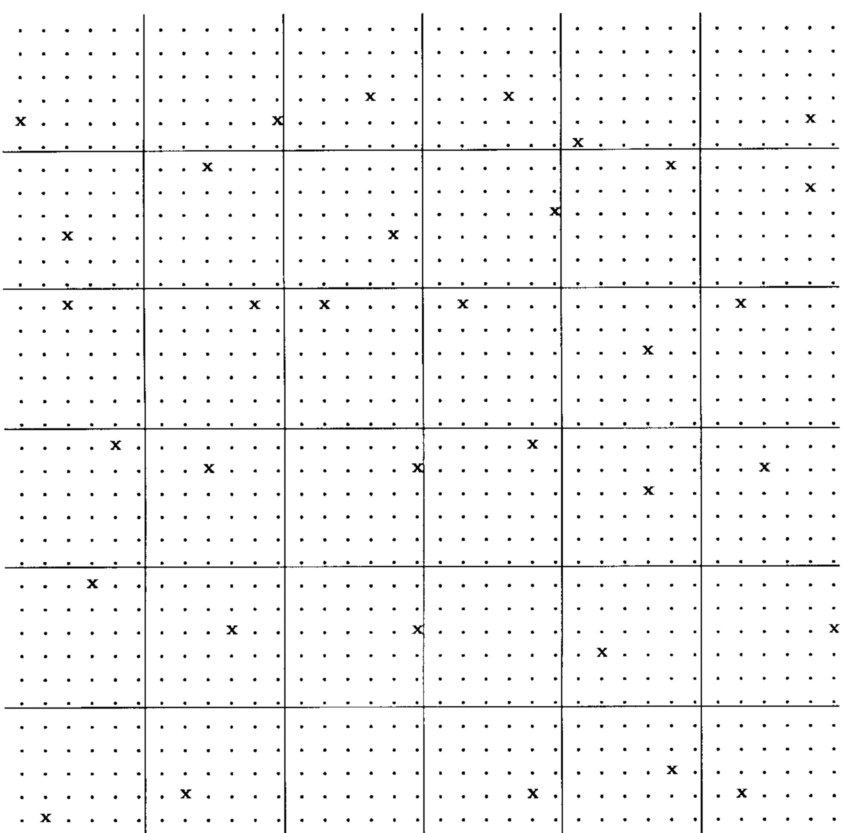

Figure 3. Spatial pattern of a random sample implemented within geographic strata, one unit per stratum. x = sampled pixel; · = unsampled pixel.

stratum is selected, an unbiased estimator of variance is unavailable (C4) and an approximate variance estimator will be necessary (see Cochran 1977, §5A.12).

5.4. *Cluster sampling*

Cluster sampling employs two sizes of sampling units, the primary sampling unit (*psu*), which is the cluster itself, and the secondary sampling unit (*ssu*), which is the subunit within the *psu*. Several options are available for defining the *psu*. Commonly the *psu* is a block of pixels, for example a 3×3 or 5×5 block, but the *psu* may be a linear arrangement of, say, 10 pixels (Edwards *et al.* 1998), an aerial photo (Zhu *et al.* 1996), a flight line along which videography has been obtained, or a mapped land-cover polygon. The *ssu* is often a pixel, but it could also be a polygon within an aerial photo or video frame. In one-stage cluster sampling, all *ssu*s within each sampled *psu* are selected into the sample. The *psu*s may be selected via SRS or a systematic design, or the *psu*s may be arranged in geographic strata and a stratified random sample of *psu*s selected. In two-stage cluster sampling, a sample of *ssu*s is obtained from each sampled *psu*. The *ssu*s within each sampled *psu* may be selected by either simple random or systematic sampling.

In cluster sampling, the assessment should be based on the correspondence between the map and reference classifications at the spatial scale of the *ssu*, not the *psu*. If the *psu* is defined to be a 3×3 block of pixels and the map and reference classifications are determined by a majority rule applied to the entire block (*psu*), for example, the spatial scale of the assessment is the *psu*. A block of pixels treated in this fashion is more appropriately viewed as a maplet (Stoms 1996) or areal sampling unit, and the approach constitutes a non-site-specific accuracy assessment at the spatial scale of the identified block of pixels. A site-specific accuracy assessment based on cluster sampling requires that a reference classification be obtained for each sampled *ssu* in the cluster, and the comparison between the map and reference land-cover classification must be on a per *ssu* basis.

The primary motivation for cluster sampling is the potential reduction in sampling costs (C6). The spatial proximity of *ssu*s within each sampled *psu* allows a larger sample size (in terms of *ssu*s) to be achieved for a fixed cost relative to simple random or systematic sampling. For cluster sampling to be effective, the *ssu*s within each *psu* should be heterogeneous; that is, neighbouring *ssu*s (pixels) within a *psu* should have different y values, where $y_i=1$ if the pixel is correctly classified, and $y_i=0$ if the pixel is incorrectly classified. This is the 'clustering principle' described by Stuart (1984, pp. 64–66). Extreme within-cluster homogeneity occurs if pixels within a *psu* are either all correctly classified, or all incorrectly classified. Within-cluster homogeneity translates into a strong positive intracluster correlation. In practice, the cost advantage of cluster sampling is diminished by the likely presence of a positive intracluster correlation, attributable to the spatial correlation of the classification errors, which leads to higher variance (Cochran 1977, §9.4). The issue then is whether the cluster sampling costs are sufficiently reduced relative to SRS to compensate for the loss of information per *ssu* attributable to the positive intracluster correlation. Moisen *et al.* (1994) provide numerical results illustrating the trade-off between sampling costs and intracluster correlation.

Two-stage cluster sampling represents a compromise between the strong spatial clustering of sampled *ssu*s in a one-stage cluster sample, and the weaker spatial clustering in the sampling units of a simple random or systematic sample. By subsampling the *ssu*s within each sampled *psu*, two-stage cluster sampling sacrifices some of the travel cost advantage (C6) of one-stage cluster sampling to gain precision (C3) by diminishing the effect of a positive intracluster correlation. Two-stage cluster sampling increases the number of *psu*s sampled relative to one-stage cluster sampling, but fewer *ssu*s are affordable at the second stage because of the increased sampling cost incurred by the larger sample size of *psu*s for two-stage cluster sampling (e.g. higher travel costs for getting to more *psu*s). Because *ssu*s from different *psu*s are expected to exhibit lower correlation than *ssu*s within the same *psu*, two-stage cluster sampling diminishes the effect of positive intracluster correlation by spreading the sampled *ssu*s among more *psu*s to reduce variance compared to one-stage cluster sampling. The comparison of one-stage to two-stage cluster sampling depends on the relative cost of sampling a *psu* versus an *ssu* and the intracluster correlation.

The cost savings achieved by the spatial control permitted by two-stage cluster sampling is strong motivation for this design, particularly for assessments covering large areas. For example, if the *psu* is an aerial photo, a first-stage sample of 50 *psu*s spatially constrains the sample to just these 50 aerial photos so that sampling costs attributable to either travel or aerial photograph expenses are reduced. Most other sampling designs, with the exception of geographic stratification by proximity to a road, do not permit such strong spatial control over the sample.

Because it is an equal probability design, one-stage cluster sampling estimation formulas for different accuracy parameters are no more difficult than those for simple random or systematic sampling (C2). Cluster sampling variance estimation formulas (C4) are more complicated than those of SRS (Stehman 1997a, Edwards *et al.* 1998) and less familiar to users. Two-stage cluster sampling may introduce unequal inclusion probabilities for the *ssu*s. If all *psu*s are the same size (e.g. the *psu* is a 5×5 block of pixels) but the sample size within *psu*s differs, then the pixels from different *psu*s have unequal inclusion probabilities. Similarly, if the *psu*s differ in size and SRS is employed at both stages, the pixels in different *psu*s will be sampled with unequal probability unless the second-stage sample size in each *psu* is proportional to the total number of pixels in the *psu*. For example, suppose the *psu* is a mapped polygon of contiguous pixels of homogeneous land-cover, and the *ssu* is a pixel. The polygons (*psu*s) will probably differ in size, so a simple random sample of, say, three pixels (*ssu*s) from each sampled polygon will result in an unequal probability sampling design. If the *ssu*s are sampled with unequal probability, estimating the accuracy parameters and associated standard errors is more complex (C2). To maintain analytic simplicity in two-stage cluster sampling, *psu*s should be equal in size and the sample size should be the same within each sampled *psu*.

5.5. *General comments on sampling design*

Choosing a sampling design for accuracy assessment requires that project objectives be clearly specified, although some flexibility in planning must be retained to accommodate unanticipated objectives that might arise once the project is underway. Planning should focus on the relative importance of the design criteria as determined by these objectives. Because sampling inference depends on the probability sampling criterion (C1), non-probability sampling designs should be ruled out. The design chosen from among the available probability sampling designs then depends on how well each design satisfies the priorities specified among the other criteria (C2–C6).

The simplicity criterion (C2) is important because if the design is implemented incorrectly, the inferences from the resulting sample will be flawed, and taking a new sample is rarely a viable option. SRS is the simplest design to augment or reduce if objectives change. Design modifications are still possible, but are more complex if strata and/or clusters are present. If multiple users and objectives are anticipated, simplicity in the analysis becomes critical. Statistical software packages may be applicable to estimate accuracy measures for some sampling designs if the proper sample weighting scheme can be implemented. It is less likely that standard errors will be computed correctly except when the design is SRS because features such as strata and clusters are generally not easily accommodated in the routine procedures of statistical packages. Providing users with a specialized analysis program to accompany the accuracy assessment data is another solution to the problem of analytic complexity (see Williams and Beach 1995). Cost savings realized by a more complex, efficient design may be diminished by the added cost of developing and overseeing a more complex analysis.

When considering the variance criterion (C3), any knowledge available on the spatial pattern of classification errors contributes to a more informed design choice. Some degree of positive spatial correlation in the errors is likely to be present, and this will generally favour systematic sampling over simple random and cluster sampling. If systematic sampling is under consideration, the expected spatial pattern of errors should be evaluated for possible periodicity, and if periodicity is strongly suspected, the systematic sampling interval should be chosen to avoid the periodicity. Statistical simultion studies may be used to provide guidance on the variance criterion (e.g. Moisen *et al.* 1994, Stehman 1992). Although simulation studies are useful to illustrate and quantify differences in precision, the results must be interpreted recognizing that population spatial patterns of error can be constructed to favour any particular design. Simulation results are valuable to the degree that the populations evaluated reflect characteristics of the real-world population that is the focus of the assessment.

When deciding whether to use either of the two population partitioning structures, strata and clusters, it is important to recognize that we are neither forced into nor prevented from using naturally occurring or existing structures to form the strata or clusters. For example, ecologically meaningful partitions such as ecoregions or forest types, partitions related to the mapping process or reference data collection such as TM scenes, flight lines or video frames, and administrative partitions such as counties or states may be considered as strata, but the option also exists to ignore these structures in the sampling design. Polygons of homogeneous land-cover may represent naturally occurring clusters, but again the sampling design need not incorporate this structure. Strata and clusters may also be imposed on a population. For example, 5×5 pixel blocks might be defined as clusters, or arbitrarily defined, but equal-sized geographic regions may be defined as strata. Such structures, even when not naturally occurring, may be advantageous to the sampling design.

The ability to improve precision by incorporating auxiliary information into the estimation component of a sampling strategy should not be overlooked. A good example of this is when the objective is to estimate overall accuracy (e.g. proportion of area correctly classified). A stratified design may be more efficient than SRS when the usual SRS estimator is used, but if SRS is combined with a poststratified estimator, this strategy becomes nearly as efficient as a proportionally allocated stratified design unless the sample size is small (Särndal *et al.* 1992, p. 267). Because an optimally allocated stratified design is usually only slightly more precise than a proportionally allocated stratified design (Cochran 1977, p. 109), a post-stratified estimator combined with SRS may perform nearly as well as an optimally allocated

Table 1 Properties of common probability sampling designs used in accuracy assessment.

Design	Analytic simplicity	Variance	Variance estimation	Spatial distribution	Spatial control
Simple random	High	High, unless combined with more complex estimation scheme	Simple	Poor	None
Systematic	High[a]	Low, unless sampling in phase with periodicity	Simple, but biased	Very good	None
Stratified					
By land cover (equal allocation)	Moderate[b]	Low for stratum-specific estimates, moderate to high for overall estimates	Moderate	Poor[c]	None
By geography	Moderate[b]	Moderate	Moderate	Very good	Low[d]
Cluster					
One-stage	High	High, unless spatial correlation is weak	Difficult	Poor to moderate[e]	Moderate
Two-stage	High[f]	Moderate, even if spatial correlation is weak	Difficult	Moderate to good[g]	Very high

[a] Systematic sampling of polygons results in unequal inclusion probabilities and a more complex analysis.
[b] High simplicity if proportional allocation is used.
[c] Good if the land-cover classes are strongly associated with geographic regions.
[d] High if unequal allocation is used to focus sampling effort in only a few strata.
[e] Poor if *psu*s are selected via SRS, moderate if *psu*s are selected via systematic sampling or by geographic stratification.
[f] Difficult if unequal inclusion probabilities at second stage.
[g] Moderate if *psu*s are selected via SRS, good if *psu*s are selected via systematic sampling or by geographic stratification.

stratified design for estimating overall accuracy. Recognizing that, a strategy combining SRS with an efficient estimator such as the post-stratified or regression estimator expands the candidate sampling strategies beyond those relying on just employing a more complex sampling design.

6. Summary

The variety of sampling designs available for accuracy assessment may at first represent a bewildering array of options. A structured approach to sample design planning helps clarify the relevant choices. The first steps are to define the population for which the accuracy assessment is needed, and to determine if this population will be partitioned into pixels, polygons, or some other areal unit. Then a probability sampling design (C1) forms the statistical foundation of the assessment. Many probability sampling design options are available for either pixels or polygons, including simple random and systematic sampling, simple random or systematic sampling applied to a population with clusters and/or strata, and one- and two-stage versions of cluster sampling. More complex designs should only be considered if the expertise is available to ensure that the design will be implemented and analysed correctly. Choosing a design from among the basic probability sampling options should be guided by project objectives, and the relative importance of the remaining five design criteria (C2 through C6). Table 1 summarizes these design criteria for the probability sampling designs reviewed in this article.

Whatever design is selected, it must be described clearly. Hammond and Verbyla (1996) have documented the lack of clarity in the description of accuracy assessment sampling designs. The description should include the population assessed, whether the population is partitioned into pixels, polygons, or some other areal unit, and the sampling design implemented. If the sampling design is not one of the classical sampling designs described in Cochran (1977) or another sampling text, the inclusion probability formulas or a means for computing the inclusion probabilities must be provided. Once the sample has been selected and the data collected, any deviations from the planned design protocol should be documented. For example, intentional exclusions of some parts of the population, missing data due to denied access or field sampling error, and any *ad hoc* methods substituted for proper design protocol should be reported.

If the sampling design for accuracy assessment is poorly chosen or improperly implemented, the scientific basis of the assessment is only cosmetically better than the 'it looks good' approach to accuracy assessment Congalton (1991, p. 45) rightly argued was unacceptable. The guidelines and principles reviewed in this article provide a basis for choosing a sampling design that provides a cost-effective, scientifically defensible assessment of map accuracy.

Acknowledgments

I thank the anonymous reviewers for their comments. Discussions with Ray Czaplewski on accuracy assessment and Scott Overton on sampling contributed much to the ideas expressed in this article.

References

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

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

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

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

Czaplewski, R. L., 1994, Variance approximations for assessments of classification accuracy. Research Paper RM-316, US Department of Agriculture, Forest Service, Rocky Mountain Forest and Range Experiment Station, Fort Collins, CO, USA..

Czaplewski, R. L., 1998, Assessment of classification accuracy and extent estimates for a land cover map. *In Spatial Accuracy in Natural Resource Analysis*, edited by H. T. Mowrer and R. G. Congalton (Ann Arbor, MI: Ann Arbor Press) (in press).

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

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

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

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

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

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

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

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

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

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

Stehman, S. V., 1996b, Sampling design and analysis issues for thematic map accuracy assessment. *Proceedings of the 1996 ACSM/ASPRS Annual Convention*, ASPRS Technical Papers, vol. 1, pp. 372–380.

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

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

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

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

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

Stuart, A. (1984). *Basic Ideas of Scientific Sampling*, 3rd edn (London: Griffin).

Williams, M. T., and Beach, D. J. C., 1995, ACAS 0.4: accuracy assessment system program manual. US Department of Agriculture, Forest Service, Rocky Mountain Forest and Range Experiment Station, Fort Collins, CO, 33pp.+source code.

Zhu, Z., Ohlen, D. V., Czaplewski, R. L., and Burgan, R. E., 1996, Alternative method to validate the seasonal land cover regions of the conterminous United States. In *Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, edited by H. T. Mowrer, R. L. Czaplewski, and R. H. Hamre, General Technical Report RM-GTR-277 (Fort Collins, CO: USDA Forest Service), pp. 409–418.