Statistical Rigor and Practical Utility in Thematic Map Accuracy Assessment

Stephen V. Stehman

Abstract

Although statistical rigor and practical utility have been advocated as desirable features of map accuracy assessment protocols, specific criteria defining these features have not been elucidated. Two criteria are proposed for statistical rigor: probability sampling and consistent estimation. Practical utility is synonymous with cost, and because cost is directly related to quality, decisions regarding practical utility may be evaluated in terms of their effect on quality. Four criteria are proposed to define quality: the precision of the accuracy estimates, the population to which sampling inference is justified, the assumptions needed to justify inference, and the accuracy of the reference data. The first step in planning a statistically rigorous, practical accuracy assessment is to construct an efficient, probability-sampling-based strategy permitting inference to the full map population. Modifications of this strategy to enhance practical utility (i.e., reduce cost of the assessment) should be evaluated using the criteria defined for quality and statistical rigor.

Introduction

Quantifying map accuracy provides important descriptive information to assess the utility of a map for a specified application. This article focuses on site-specific, thematic accuracy (Janssen and van der Wel, 1994; Stehman and Czaplewski, 1998; Congalton and Green, 1999) in which accuracy is defined by comparing the map attribute and the actual attribute (i.e., reference classification) for a sample of pixels, polygons, or other areal units such as a 1-hectare plot. Statistical inference is then applied to generalize or extrapolate the results from the sample to the full map population. Most accuracy assessment objectives can be addressed by design-based inference (Stehman, 2000), the inferential framework typically invoked in classical sampling theory and methods (Cochran, 1977; Sarndal *et al.*, 1992, p. 515). Design-based inference is assumed throughout the remainder of this article.

Because budget constraints affect nearly all accuracy assessment projects, cost becomes a dominant concern in planning. What can be done to make the assessment more cost-effective, and therefore more practical, while still maintaining statistical rigor? The proposed approach is to first construct a statistically rigorous, efficient sampling strategy employing traditional concepts and methods of scientific sampling. If it is then necessary to sacrifice some desirable features of the assessment protocol to reduce costs, the options include reducing precision, restricting the population to which design-based inference applies, introducing assumptions, and allowing greater error in the reference data. Implementing one of these options should be done only with full awareness of its effect on the statistical rigor of the assessment.

A theme of this article is that both statistical rigor and practical utility can be evaluated using very specific criteria. These criteria are described, and it is demonstrated how they can be applied when planning an accuracy assessment. The first part of the article addresses the concepts required to define a statistically rigorous assessment. The second part of the article discusses the effect on the quality of an accuracy assessment when various cost-reduction measures are taken.

Components of Accuracy Assessment

Stehman and Czaplewski (1998) list three basic components to accuracy assessment: the response design, sampling design, and analysis. The response design is the protocol for determining the reference classification recorded on each sampling unit. Sometimes a single attribute, for example, the primary land-cover class, is recorded. Alternatively, the response design may specify recording the primary and any secondary land-cover classes, a value for each land-cover class derived from a linguistic scale (Gopal and Woodcock, 1994), or the proportion of each land-cover class present in the assessment unit. Choosing the spatial unit on which the assessment is based is also part of the response design protocol. The spatial unit may be a pixel, block of pixels, land-cover polygon, or other areal unit. The criteria defined for statistical rigor and practical utility and the basic sampling theory and concepts discussed throughout this article apply to any of these choices of assessment unit.

A response design is present in all sampling problems. Determining the reference land-cover classification is analogous, for example, to measuring intelligence, common sense, or income when sampling human populations. Each of these attributes of a human population must be defined, and scientists may not agree on a definition or measurement protocol for a given attribute. Some human attributes are easier to measure than others, just as some land-cover types are more readily identified than others. Whether the focus is on attributes of humans or landscapes, the response design represents a measurement problem, not a statistical inference problem. Although reference data quality is important to accuracy assessment, it is useful to separate measurement issues of the response design from the inference issues affecting sampling design and analysis.

The sampling design is the protocol by which the sample elements are selected. Stehman and Czaplewski (1998), Congalton and Green (1999), and Stehman (1999) review options for the sampling unit and sampling design. Analysis typically focuses on estimating an error matrix (Story and Congalton, 1986) and various measures summarizing the error matrix.

SUNY College of Environmental Science & Forestry, 320 Bray Hall, Syracuse, NY 13210 (svstehma@mailbox.syr.edu).

Photogrammetric Engineering & Remote Sensing
Vol. 67, No. 6, June 2001, pp. 727–734.

0099-1112/01/6706–727$3.00/0

Congalton (1991), Janssen and van der Wel (1994), and Stehman (1997) review common summary measures applied in accuracy assessment.

Proposed Criteria for Statistical Rigor

The sampling design and analysis components are directly linked to statistical inference and therefore motivate the two criteria proposed for statistical rigor. A statistically rigorous accuracy assessment is one in which the sampling design satisfies probability sampling protocol and the estimates are statistically consistent. Both criteria will be described in detail, and then a basic result of sampling theory will be presented to show how these two criteria are unified within a general framework of estimation.

A probability sampling protocol is one in which the inclusion probabilities are known for all elements in the sample, and the inclusion probabilities are non-zero for all elements of the population. The inclusion probability for element u, denoted π_u, is defined as the probability that element u is included in the sample (u may denote a pixel, polygon, or other sampling unit chosen for the assessment). Sarndal *et al.* (1992, Section 2.4) discuss inclusion probabilities and probability sampling from a general perspective, and Stehman and Czaplewski (1998), Biging *et al.* (1998), and Stehman (1999) discuss probability sampling in the context of accuracy assessment.

Inclusion probabilities affect two key practical elements of accuracy assessment: they determine unambiguously (i.e., without assumption) the population represented by the sample, and they are required for design-based estimation. Population representation derives from the requirement that the π_u's must be non-zero. The population to which rigorous statistical inference applies is then the collection of all elements for which $\pi_u > 0$. Elements of the population for which $\pi_u = 0$ are effectively excluded from being sampled, and design-based inference does not extend to those elements. For example, restricting sampling to polygon interiors excludes a portion of the population from any chance of being sampled, and inference would not apply to polygon edges.

Requiring the π_u's to be known for the sample is necessary for estimation. A basic theorem of probability sampling (Horvitz and Thompson, 1952) is that if $\pi_u > 0$ for all elements in the population, then the Horvitz-Thompson (HT) estimator $\sum_s y_u/\pi_u$ is unbiased for the population total $\sum_U y_u$ ($\sum_s$ indicates summation over the sample elements and $\sum_U$ denotes summation over the population). The HT estimator is applied frequently in sampling practice, although usually in more convenient special case forms available for the standard sampling designs often employed in practice (Sarndal *et al.*, 1992, Secs. 2.8, 3.4.1, 3.7.2, 4.2.1). Typically, in accuracy assessment, $y_u = 1$ if element u is classified correctly, and $y_u = 0$ otherwise. Most accuracy parameters may be formulated as totals, and therefore can be estimated by the HT estimator or functions of HT estimators.

The HT estimator is introduced because it provides insight into the conceptual basis of rigorous estimation. Writing the estimator as $\sum_s w_u y_u$, where $w_u = 1/\pi_u$ highlights the necessity to weight each sample observation to expand its representation to account for elements of the population not sampled. Stuart (1984) coined the term "apparent frequency" for this weight or expansion factor, w_u.

A simple example illustrates the concept. Suppose we would like to estimate the total amount of money for a population of $N = 100$ people. A simple random sample (SRS) of $n = 10$ people is selected, and each sampled person reports the amount of money, y_u, he or she has. To expand the sample data to the full population of 100 people, each sampled person must represent a certain number of people in the population. For the SRS protocol, we would intuitively expect each sampled person to represent himself or herself plus nine others. Formally, because each person has a probability of $\pi_u = 1/10$ of being included in the sample, each sampled person must represent $w_u = 1/\pi_u = 10$ people in the population. A sampled person having, say, \$2.60, would be assumed to represent $w_u y_u = (10)(2.60) = \$26$ for the population. If we did not select the sample using a probability sampling protocol, we cannot operate within the design-based estimation framework because the probabilistic basis for expanding from the sample observations to the population has been lost.

Different sampling designs lead to different π_u's and, consequently, different weights. In the illustrative example, if the population is stratified by gender, and male and female strata are sampled with different intensity, the expansion weight for sampled males will differ from the expansion weight applied to females. This weighting feature applies directly to estimation in accuracy assessment. Each sampled pixel, polygon, or other areal unit must represent a certain number of these units in the population. The sample units do not all need to have the same weight, but these weights must be known for each sample unit.

Consistent Estimation

The estimation criterion proposed for statistical rigor is consistent, not unbiased, estimation. Sarndal *et al.* (1992, Sec. 5.3) provide a technical definition of consistency. Heuristically, we will not go far wrong in thinking of consistent as synonymous with unbiased: the desirable feature is that the sampling distribution of the estimator should be centered on the target parameter to be estimated. Consistency encompasses situations for which an unbiased estimator is problematic, as for example when the accuracy parameter is a ratio of two or more totals.

Two results complete the theory required for rigorous estimation in accuracy assessment: (1) the HT estimator is consistent for a population total, and (2) a continuous function of HT estimators is consistent for the parameter defined by the corresponding function of the population totals (Sarndal *et al.*, 1992, Remark 5.3.1). To derive a consistent estimator, we formulate the accuracy parameter as a function of population totals, and then estimate each total using HT estimation. Stehman (1996b) illustrates this technique to derive an estimator of kappa for stratified random sampling. Because accuracy assessments typically rely on standard sampling designs, it is often unnecessary to derive estimators from these basic principles. However, this approach to constructing consistent estimators is a fundamental structure underlying a statistically rigorous assessment.

Assumptions

The design-based estimation theory applied to accuracy assessment is remarkably free of assumptions. The consistency and unbiased properties of Horvitz-Thompson estimation do not require any assumptions concerning the probability distribution, independence, or variance of the observations (Horvitz and Thompson, 1952; Sarndal *et al.*, 1992). Recognizing when assumptions are not necessary is as important as knowing when assumptions are required to justify inference. Consider the statement: "The concept of randomness is a central issue when performing almost any statistical analysis because a *random sample is one in which each member of the population has an equal and independent chance of being selected*" [italics added] (Congalton and Green, 1999, p. 24). Two unnecessary restrictions—equal probability and independence—are implied by this statement. Although many of the hypothesis testing methods familiar from introductory statistics require independence and equal probability, the descriptive analyses comprising a majority of accuracy assessment estimation objectives do not. The design-based estimation formulas for overall accuracy, user's and producer's accuracies, and the cell proportions of the error matrix are unbiased and/or consistent without the need for these assumptions. Unnecessarily invoking the independence restriction remains prevalent in sampling practice (Gregoire, 1998; Stehman, 2000). Misplaced concern over unnecessary assumptions may distract practitioners from the relevant criteria defining statistical rigor that should be the focus in planning an accuracy assessment.

Practical Implications of the Statistical Rigor Criteria

It is useful to translate the two criteria proposed for statistical rigor into practical guidelines. If a standard sampling design such as simple random, stratified random, systematic, or cluster sampling is correctly implemented, the probability sampling criterion is satisfied. Conversely, the probability sampling criterion is violated when the sample units are selected by judgment to be "representative," selected because of convenient access (e.g., near roads, schools, or on public land), or selected because of homogeneity of land cover (e.g., polygon interiors). Rigorous design-based sampling inference is not possible from such non-probability sampling protocols. Hammond and Verbyla (1996) note the potential optimistic bias inherent in restricting sampling to polygon interiors or other homogeneous areas of land cover. Paulsen *et al.* (1998) and Peterson *et al.* (1999) provide numerical documentation of the potential impact non-probability sampling may have in practice. They found that estimates obtained from a non-probability sample of conveniently accessed lakes corresponded poorly with estimates derived from a probability sample of lakes. It would not be surprising to find correspondingly poor estimates from non-probability samples employed in accuracy assessment.

The probability sampling criterion is also violated when the selection protocol is so complex that it is impossible to determine the π_u's. Congalton and Green (1999, Chapter 8) present an illustration of this difficulty. In their example, they describe several steps taken in the sampling protocol to diminish problems arising from access to private property, travel costs, and distance to a road. Although randomization is incorporated in this protocol, deriving the π_u's poses a daunting challenge. The practical implication is that, if these π_u's cannot be specified, the data are unusable for rigorous design-based inference. Stehman and Czaplewski (1998) recommend that, if the π_u's resulting from a selection protocol cannot be determined, the protocol should be replaced by an alternative sampling design for which the π_u's are known.

Most accuracy assessments will be subject to practical difficulties affecting implementation of the design. For example, Congalton and Green (1999, p. 94) report that approximately 50 percent of their randomly selected sample locations could not be visited in the field, and Edwards *et al.* (1998, p. 80) document several practical difficulties of visiting ground locations. The remedies to these practical problems must be based on sound sampling principles. *Ad hoc* modifications may prove unacceptable because it is often very difficult to derive π_u's for the field-based decision protocols often implemented.

Consider the following example in which the sampling unit is a mapped polygon and the design is simple random sampling (SRS) from a list frame of all map polygons. Rather than visit just these sample polygons, it is decided that a convenient way to increase sample size is to instruct field crews to sample also the nearest polygon due east of the original sample polygon. A similar protocol of going to the nearest accessible pixel or polygon (perhaps of the same land-cover class) is sometimes recommended to replace denied or difficult-to-access sites. The π_u's created by this modification are now changed and drastically more complicated to derive than the π_u's of the original SRS design, which are all equal. In the modified protocol, selecting the nearest polygon to the original sampled polygon creates an unequal probability sampling design because polygon size and spatial arrangement determine the π_u's. The π_u's increase as a function of polygon area (i.e., larger polygons are more likely to be selected), but translating this relationship into correct formulation of the π_u's will be difficult. Further, it may be impractical to obtain the field measurements required to work out the geometry necessary to compute the π_u's. The desire to decrease travel costs by visiting polygons in close proximity is justified. But the purpose of this example is to emphasize that it is critical to achieve this practical goal while still maintaining the statistical rigor provided by probability sampling. In most situations, a sampling technique exists to achieve the dual purpose of statistical rigor and practical utility. Two-stage cluster sampling would apply to this particular situation.

If the sampling protocol is such that the π_u's are known, the consistency criterion of statistical rigor is readily satisfied in practice by incorporating the π_u's in the accuracy estimators. For equal probability designs such as simple random, systematic, and some forms of cluster sampling, each sample observation receives the same weight and estimation formulas typically presented for accuracy assessment (Janssen and van der Wel, 1994; Congalton and Green, 1999) are appropriate. Equal probability sampling designs are called "self-weighting" because, by construction of the design, all sampling units have the same weight. If the design includes unequal probability sampling, for example, stratified sampling with equal allocation, then the weights determined by the design must be incorporated in the estimation. To apply SRS estimation formulas to an unequal probability sampling design violates the consistency criterion and will typically result in unacceptable estimates.

Practical Utility: Sampling Design and Analysis

Practical utility is viewed as a function of cost. With few exceptions, any practical problem can be greatly diminished or eliminated if no expense is spared. Unfortunately, the reality of accuracy assessment budgets dictates that cost is a crucial, if not dominant, planning consideration. To reduce costs, an efficient sampling strategy should be constructed to achieve the best precision possible given the available resources. The next subsection discusses elements of an efficient sampling strategy. If further cost reductions are necessary, even after an efficient sampling strategy has been chosen, lowering costs by reducing quality may be considered. Four general criteria of quality are proposed and discussed in later subsections.

Designing an Efficient Sampling Strategy

The first step in developing a cost-effective accuracy assessment protocol is to construct an efficient sampling strategy. For most practical problems encountered in accuracy assessment, a statistically sound sampling procedure exists. Large, spatially extensive populations, multiple estimation objectives, and difficult to access sampling units are not characteristics unique to sampling design for accuracy assessment, but are problems which have existed throughout the history of sampling practice. A diverse collection of scientifically sound methods has been assembled (Hansen *et al.*, 1953; Kish, 1965; Murthy, 1967; Raj, 1968; Cochran, 1977; Jessen, 1978; Thompson, 1992; Schreuder *et al.*, 1993). Productive advances in accuracy assessment may derive from innovative application of these traditional sampling methods. Both the sampling design and analysis components of a sampling strategy offer options for improving cost-effectiveness. The planning goal should be to select the sampling strategy that most efficiently uses the budgeted resources. A few examples will be described to illustrate how existing methods address common practical problems in accuracy assessment.

Stratified sampling provides for efficient estimation within population subgroups. Frequently, the strata are the mapped land-cover classes chosen for the objective of estimating user's accuracies. But geographic stratification based on ecoregions or spatial regions of special interest is another option. Stratification may be used to ensure a minimum sample size in each stratum to achieve more precise (i.e., efficient) estimates for rarer land-cover types or smaller subregions than would be obtained from simple random or systematic sampling. Stratification may also be employed to reduce sampling costs, for example, by stratifying by distance to a road or by ease of accessibility (Edwards *et al.*, 1998). Optimal allocation formulas (Cochran, 1977, Sec. 5.12) can be used to determine the most efficient allocation of sampling resources to strata when the objective is estimating overall accuracy. Sarndal *et al.* (1992, Sec. 12.7) present formulas for efficiently allocating samples among strata for multiple estimation objectives (e.g., user's accuracies and overall accuracy).

Cluster sampling using a county, USGS quadrangle, aerial photograph, or another areal unit as the primary sampling unit (PSU) is often adopted to reduce travel costs. Cluster sampling is also applicable to other situations in which spatial properties of the sample affect the practical utility of implementing the design. For example, if identifying and contacting land owners is expected to increase costs dramatically, using a county as a cluster may reduce the cost of identifying property ownership as well as decrease travel costs to sample sites. Defining counties as clusters creates a design option in which having to sample elements in every single county of a state or region would not be necessary, thus reducing the number of courthouse visits needed to determine ownership.

A potential disadvantage of cluster sampling is that classification errors tend to cluster spatially, so that the information per sample unit in cluster sampling may be lower than for other designs. This problem is diminished by two-stage cluster sampling in which a subsample of the elements within a PSU are sampled. For example, Edwards *et al.* (1998) employed USGS 7.5-minute quadrangles as PSUs, and Zhu *et al.* (2000) constructed PSUs based on NAPP photography. The spatial restriction of the sample to within the first-stage PSUs reduces travel costs if ground visits are necessary, or reduces the workload for obtaining aerial photography or videography to only those areas covered by the first-stage sample. Czaplewski (1999) describes general multivariate statistical estimators applicable to analysis of data obtained from two-stage cluster sampling.

A variation of cluster sampling—adaptive cluster sampling (Thompson, 1990)—has been proposed for sampling rare land-cover classes (Stehman, 1996c) and for change-detection accuracy assessment (Biging *et al.*, 1998). A potential advantage of adaptive cluster sampling is to increase the sample size from rare land-cover classes as identified by the reference rather than the mapped classification. It is simple to capture rare mapped classes in the sample by stratifying by the known mapped land-cover classes, but if these rare classes are poorly mapped, we may get few true elements of this rare class in the sample. The practical utility of adaptive cluster sampling has not been directly confirmed in accuracy assessment applications, but it offers a potentially useful, cost-effective option meriting consideration.

Estimator precision can be improved at the analysis stage by using poststratified (Card, 1982) and regression estimation (Stehman, 1996a; Kalkhan *et al.*, 1998; Czaplewski, 1999). These techniques require no additional field visits. Poststratification, which incorporates the mapped land-cover proportions into the estimator, is always an option because these mapped proportions are known. Regression estimation requires some auxiliary information related to accuracy, such as photointerpreted land-cover labels when the reference label is based on ground visits. These estimation techniques increase complexity of the analysis, but usually incur little or no additional sampling cost.

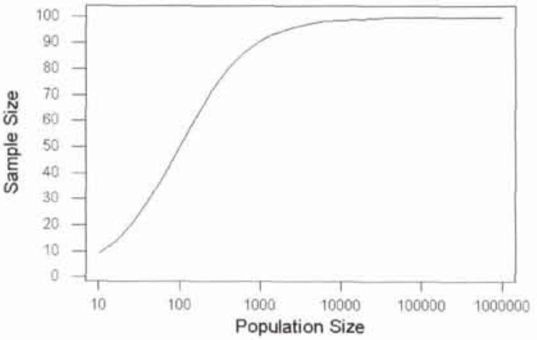

Figure 1. Sample size required to achieve SE$(\hat{P})$ = 0.05 for different population sizes.

Relationship between Sample Size and Population Size

Although classical sampling techniques have historically played an important role in providing scientifically defensible estimates for large populations (Bellhouse, 1988), the applicability of "traditional thinking about sampling" to accuracy assessment has been questioned because of the large population sizes (e.g., number of pixels) common in remote sensing practice (Congalton, 1991, p. 43; Jensen, 1996, p. 249; Congalton and Green, 1999, p. 18). Traditional sampling techniques do, in fact, apply to these large populations because precision is determined by the absolute sample size, n, not the size of the sample relative to the population size. Large populations do not require larger absolute sample sizes to obtain precise estimates, so it is not necessary to sample the same fixed percent of a large population as would be required to attain similar precision for a smaller population.

These relationships are illustrated in Figure 1 in which the sample size required to achieve a standard error of 0.05 for the estimated proportion of correct classifications, $\hat{P}$, is shown as a function of population size, N (population size is plotted on a logarithmic scale, $\log_{10}(N)$, so that very large population sizes can be accommodated in the figure). Suppose the true accuracy is P = 0.5 and the design is SRS. The standard error of the estimated proportion is then SE $(\hat{P}) = [(1 - n/N)P(1 - P)/n]^{1/2}$. If the precision objective is to obtain a standard error of 0.05, solving for n yields $n = N/(0.01N + 1)$.

Figure 1 shows that the sample size required to achieve a standard error of 0.05 initially increases as a function of N, but then reaches a plateau of $n = 100$ at about $N = 10{,}000$. For $N = \infty$, the required sample size is $n = 100$, thus identifying the upper bound on n. No matter how many pixels or polygons are present in the mapped region, a sample size of 100 will ensure that P can be estimated with a standard error of no greater than 0.05.

When the sample size required to achieve a standard error of 0.05 is expressed as a percent of the population size, (n/N)*100%, the percent of the population that must be sampled shrinks toward 0 as N increases (Figure 2). Clearly, to achieve SE $(\hat{P}) = 0.05$ for larger populations, it is not necessary to sample the same fixed percent of the population. Sampling a very small percent of a large population can yield precise estimates.

If planning is based on sampling a fixed percent of the population, the resulting sample size for a large population will be much larger than necessary to achieve the target standard error. For example, for $N = 20{,}000$ and $P = 0.5$, a sample size of $n = 100$ would yield the targeted standard error of 0.05. This

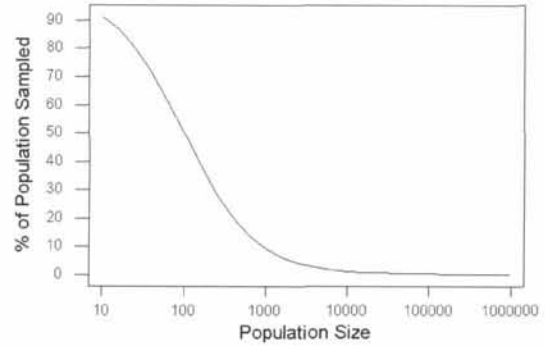

Figure 2. Percent of the population that must be sampled to achieve $\text{SE}(\hat{P}) = 0.05$ for different population sizes.

translates to a $100/20{,}000 = 0.5$ percent sample. Now suppose the population size is $N = 300{,}000$ and the same sampling intensity of 0.5 percent is applied. The resulting sample size is 1500 and the standard error is 0.013, so that the sample size based on the 0.5 percent guideline would be much larger than necessary to achieve the precision goal of 0.05. The overly precise estimates resulting from the 0.5 percent rule may be an inefficient use of limited accuracy assessment resources. A sample size of $n = 100$ yields nearly the same standard error for a population of 20,000 pixels (or polygons) as it does for a population of 300,000 pixels (or polygons).

The relationship between n and N might be characterized as "smaller populations allow smaller sample sizes to achieve a set level of precision, but larger populations do not require increasingly larger sample sizes." Cochran (1977, Sec. 2.6) and Stuart (1984, Sec. 8) provide additional clarification of this relationship. Because good sampling methodology is so fundamental to statistically defensible accuracy assessments, it is important to recognize the appropriateness, if not the necessity of applying "traditional thinking about sampling" to accuracy assessment, despite statements to the contrary.

Practical Utility: Cost versus Quality Tradeoffs

Once the cost savings of an efficient strategy have been achieved, we may consider trading quality for further cost reductions. The quality criteria or "resources" we have available to exchange for reduced cost include precision, population representation, assumptions, and accuracy of the reference data. The details of these tradeoffs are described in the next five subsections.

Precision

Smaller sample sizes cost less, and reducing sample size lowers precision (assuming the sampling strategy is unchanged). Because assessments are typically designed to estimate several accuracy measures, some flexibility exists to distribute the loss of precision among selected estimates. For example, suppose the design employs stratification by mapped land-cover class, and the goal is to estimate user's accuracy of each class with a standard error of 0.05. Equally reducing the sample size of all strata would increase the standard error of all user's accuracy estimates. Alternatively, if the land-cover classes differ in importance to the mapping objectives, greater sample size reductions could be exacted from the less important classes, thus preserving better precision for the estimates of the more important classes.

Consider another scenario in which the original objectives specify a target standard error for the estimated accuracy of each combination of an ecoregion by land-cover category cross-classification. Suppose there are five ecoregions and 20 land-cover categories, and budget limitations subsequently deter the initial ambitious estimation objective because resources are not available to achieve the precision goal for each of the 100 estimates desired. To enhance the practical utility of this assessment, we could reduce the overall sample size needed by applying the precision requirement only to the land-cover class estimates aggregated over all ecoregions. This smaller overall sample size may then be adequate because only 20 estimates must meet the precision standard. Given this smaller sample, it is still possible to estimate accuracy for any land cover by ecoregion cross-classification. But standard errors may be high for rare classes in small ecoregions, perhaps so high that estimates are not practically meaningful (e.g., estimated accuracy is 0.55 with a standard error of 0.40). In this example, practical utility (in terms of smaller sample size) is gained at the expense of poorer precision for selected estimates.

Population Representation

Lowering the cost of accuracy assessment by eliminating sites distant from roads, on private land, or in otherwise inconvenient to access locations reduces the population to which inference applies. In this subsection, it is assumed that a probability sample is obtained from the portion of the mapped area not excluded by the imposed restrictions. Consequently, rigorous design-based inference still applies, but only to that portion of the original population for which $\pi_u > 0$. For example, if probability sampling is restricted to public land, design-based inference applies only to the area represented by all public land. This population representation issue is sometimes framed by defining target and sampled populations (Cochran, 1977, p. 5). The target population is the population for which inference is desired, usually the entire mapped region in accuracy assessment. The sampled population is that population from which we have a probability sample. Statistical inference applies to the sampled population. If the sampled and target populations differ, generalization to the target population will require assumptions (see next subsection).

When population representation is reduced to save costs, it is important to describe the sampled population. Providing a map of the area represented by the sampled population or reporting the percent of the original population represented by the inference (e.g., the percent of area in public ownership) are options for describing the sampled population to which the inferences apply. Stehman and Czaplewski (1997) provide additional recommendations for describing the sampled population when sampling is restricted to a subset of the mapped population.

Assumptions

Another option to reduce costs is to replace the protocols of probability sampling and consistent estimation by assumptions. Consider again the situation in which sampling is restricted to a reduced population. In the previous subsection, inference was correspondingly restricted to the reduced (sampled) population. Now suppose inference to the full target population is desired. A simple assumption to invoke is that accuracy of the reduced population is representative of accuracy of the full population (e.g., classification accuracy for inaccessible areas is similar to classification accuracy for accessible areas). Support for such an assumption may be argued on the basis that similarity of the sampled and target populations translates into similarity of classification accuracy. Comparisons of the target and sampled populations might be based on mapped land-cover area, soils, geology, physiography, ecoregions, or any other theme available in a GIS, or also on averages or distributions for physical variables such as elevation, precipitation, or temperature. Even if these characteristics are similar between the sampled and target populations, the possibility remains that classification accuracy still differs between the two populations because of differences not captured by the variables used in the comparison (e.g., land management history and productivity).

Different assumptions are required when the π_u's are unknown. One option is to assume values for the unknown π_u's and then to proceed to estimate accuracy parameters on the basis of these assumed π_u's. A common example of this approach is to analyze the sample data as if the design had been SRS. This is the analysis implemented by Congalton and Green (1999, Chapter 8) for the data resulting from the complex selection protocol described in their example. The actual, but unknown, π_u's generated by the complex protocol are replaced by those appropriate for SRS (i.e., probability sampling protocol is replaced by an assumption of SRS). Inference derived from this implicitly assumed model of the π_u's is not well supported because the complex data collection protocol undoubtedly results in unequal π_u's. Rather than weight all sample units equally, the analysis should incorporate different weights for different sampling units. Consequently, the analysis based on SRS formulas violates the consistency criterion.

Modeling π_u's has an analogy in line transect sampling for estimating animal population abundance (Burnham *et al.*, 1980). A model is constructed to represent detectability of the animals selected in the sample, and estimation is based on the detectability model rather than on the actual, but unknown, π_u's. These wildlife applications focus on mobile animal populations for which probability sampling designs are often impractical. Because accuracy assessment problems are much more amenable to probability sampling, modeling π_u's should not be adopted as standard practice in accuracy assessment.

Implementing a sampling protocol for which the π_u's are unknown creates inferential risks. If the probability sampling criterion is not satisfied, then it is difficult to specify unambiguously the population to which inference applies. Inference becomes entirely dependent on assumptions and, as such, will be very difficult to defend in a confrontational setting.

Quality of Reference Data

A final option for reducing costs is to decrease reference data quality. For example, suppose ground visited reference data are replaced by less expensive photointerpreted reference data. For some classification schemes and in some environments, this may result in little or no deterioration of reference data quality. However, in those situations in which ground visits are deemed more accurate, a decision must be made on whether lower quality photointerpreted reference data represent an acceptable compromise. Ground visit sampling costs could be reduced by taking fewer measurements or by investing less effort in precise spatial location of the sample points. Replacing actual measurements of tree diameter, canopy cover, ground cover, or species composition, for example, by visual approximations is another cost-cutting measure. Simply employing less well-trained photointerpreters or field technicians reduces cost and, unfortunately, reference data quality.

Cost reductions in the response design require intense scrutiny because poor quality reference data diminish the quality of the accuracy assessment. Crist and Deitner (1998) describe some of the practical difficulties and cost considerations affecting the quality of reference data. Congalton and Green (1999, Chapter 4) provide an excellent detailed discussion of potential sources of error in reference data, and recommend that these errors should be quantified. Husak *et al.* (1999) suggest an interesting approach to quantify the effect of spatial registration error in the reference data. Statistical techniques for incorporating measurement error (e.g., assigning an incorrect reference label) into data analysis exist (Cochran, 1977, Chapter 13; Sarndal *et al.*, 1992, Chapter 16), but these methods have rarely been applied to accuracy assessment problems. Analyses incorporating measurement error typically require a statistical model. Therefore, the criteria proposed for rigorous design-based inference may not be sufficient to accommodate the analysis of accuracy data possessing a high degree of measurement error. Specifying criteria for rigorous inference when the analyses depend on a model would be a useful contribution to the accuracy assessment literature.

Example: Tradeoffs when Intensifying Sampling Near Roads

An example summarizes some of the tradeoffs between cost and the four criteria of quality described in the previous subsections. Suppose that the cost of ground reference samples increases with distance from a road. Cost is a primary consideration in planning, and several sampling design options, each differing in cost, are considered. Option 1 is stratified random sampling employing two strata defined by distance from a road (>1 km and ≤1 km). A larger sample size is planned for the less expensive to sample, near-road stratum. Option 2 is a probability sample of only locations within 1 km of a road. Option 3 is also initiated from a probability sample limited to locations within 1 km of a road but, to further reduce costs, a few convenient staging sites are selected at which field crews will be based. The field crews then travel to the probability sampling sites beginning with those closest to the staging sites until a specified minimum number of sites have been visited. The remaining unvisited sample sites are discarded. In Option 4, again only locations within 1 km of a road are visited, but these locations are not identified by a probability sampling protocol. Instead, the locations visited are conveniently accessed from the staging sites described in Option 3. In effect, Option 4 is a reconnaissance windshield survey. A fifth option, unrelated to sampling design, is to avoid ground visits entirely and to derive the reference data from photointerpretation.

Options 1 through 4 are ordered by decreasing cost and decreasing quality. All four options are less expensive than simple random or systematic sampling, neither of which is constrained to intensify sampling closer to roads. Option 1 is the only probability sample unambiguously representative of the full population. This option may be more expensive than the other options because it forces some sampling to occur away from roads. But rigorous inference to the full map population is justified without assumption. Option 2 is also a probability sample and therefore provides rigorous design-based inference, but only to the reduced population defined by the area of the mapped region within 1 km of a road. If Option 2 is implemented, the accuracy report should clearly specify this reduced population representation.

Option 3, being initiated from a probability sample, allows inference to the same reduced population described for Option 2. But visiting sample locations in a sequence established by distance from the staging sites introduces a complicated deviation from the original probability sampling structure of Option 2. Those locations closer to staging sites have a higher probability of being included in the sample. Deriving the π_u's for Option 3 will be problematic, thus jeopardizing the goal of satisfying the consistency criterion. It may be necessary to assume values for the π_u's, or to assume that accuracy is the same near the staging sites as it is distant from the staging sites. Neither assumption may be tenable. Option 4 is not a probability sampling design, and inferences drawn from the data are completely dependent on assumptions.

When choosing among these options, the evaluation should focus on the four specific criteria of quality proposed. Accordingly, Option 4 is clearly the worst choice, forfeiting the inferential rigor provided by probability sampling. The randomization present in Option 3 makes this option more satisfactory than Option 4, but the sampling protocol of Option 3 creates the problem that it may be impractical to satisfy the consistency criterion because of the difficulty of determining the π_u's. Option 1 has an advantage over Option 2 because it allows inference to the full target population. However, the precision criterion may influence the choice between Options 1 and 2. If sampling distant from a road is so expensive that the overall sample size is small for Option 1, precision may be unacceptably poor. Because of the lower cost of sampling near roads, Option 2 might permit a larger sample size, producing precise estimates, although for a reduced population. If the budget is very small, a difficult decision must be made: Is it better to obtain reasonably precise estimates for a reduced population (Option 2), or to estimate imprecisely parameters of the full target population (Option 1)?

Summary

Statistical rigor and practical utility are desirable features of any accuracy assessment strategy. To effectively incorporate statistical rigor and practical utility into the planning process, these features must be explicitly defined. Probability sampling and consistent estimation are proposed as the two criteria determining statistical rigor. Practical utility is framed in terms of tradeoffs between cost versus quality, where the proposed criteria of quality are estimator precision, the population represented by the inference, the assumptions needed to justify inference, and the accuracy of the reference data. The criteria for statistical rigor and practical utility are defined within the context of design-based inference, and the implications of these criteria on the practice of accuracy assessment are similarly specific to design-based inference. Different criteria and practical implications would likely result from taking a model-based inference perspective.

Practical constraints are a reality of accuracy assessment. Circumventing these practical problems is often extremely difficult and may require innovative sampling designs or sophisticated statistical analyses. But it is possible to construct a statistically rigorous, yet practical, sampling strategy for accuracy assessment by adhering to the criteria proposed in this article. Classical sampling methods provide a diverse collection of efficient techniques from which to choose, consistent estimation provides a rigorous approach for analysis, and the proposed criteria defining quality supply a basis for choosing judiciously among options for reducing costs.

The criteria defining statistical rigor have important practical implications on accuracy assessment. The protocol that produces the data from which an error matrix and other summary measures are estimated determines the appropriateness of the data for inference. All data are not equally valuable. Those data arising from a probability sampling protocol can be unambiguously associated with a population to which design-based sampling inference is justified, and the consistency criterion ensures that estimates apply to parameters of this population. Probability sampling provides the basis for the estimation procedure (using the weights derived from the π_u's) and establishes representativeness of the inference. Inferences from data obtained using a non-probability sampling protocol are on much less certain footing. Poorly designed, statistically flawed data collection methods still produce data, error matrices, and accuracy estimates, but these estimates represent quantitative window dressing to an assessment that is only cosmetically better than an "it looks good" approach. Cavalierly applying simple analyses (e.g., SRS estimation formulas) to *ad hoc* and/or complex data collection protocols will produce estimates of unknown inferential value and dubious credibility.

The statistical rigor of an accuracy assessment protocol is not apparent from the error matrix or from the accompanying accuracy estimates. Therefore, the sampling design and analysis must be reported in a manner that provides a clear depiction of the protocol. Ideally, information would also be provided to evaluate the four proposed criteria of quality. Reporting standard errors supplies the necessary information to evaluate precision (assuming that the standard errors are computed correctly for the sampling design and analysis implemented). Population representation and assumptions required for inference can be evaluated if the sampling protocol is clearly documented. For example, restrictions on the population sampled and deviations from probability sampling protocol should be noted. Any assumptions invoked in the analysis should be explicitly stated, and diagnostics evaluating the validity or effect of the assumptions should be reported if available. To evaluate reference data quality, detailed description of the response design protocol, perhaps accompanied by some of the analyses suggested by Congalton and Green (1999, Chapter 6), provides critical information.

Accuracy assessments may be viewed as falling along continua of statistical rigor and quality, depending on the objectives and user needs underlying the assessment. The highest quality, most rigorous assessments will provide precise estimates from a probability sample representing the full target population and based on very accurate reference data. If the budget is inadequate to support such an assessment, cost may be decreased by accepting less precise estimates, possibly for a reduced population of inference. Requiring rigorous inference to a known, identifiable population should be non-negotiable. If budget constraints prevent implementing a probability sampling design or if sample size is so small that estimates have poor precision, it may be more prudent to skip the assessment entirely if minimal standards of precision and statistical rigor cannot be met.

By recognizing specific criteria for statistical rigor and quality, we can evaluate the consequences of various accuracy assessment planning decisions. The selected protocol must be statistically rigorous, practical (i.e., cost-effective), and also of high quality. To achieve all three of these desirable features, criteria of statistical rigor and quality must be taken into account when planning and implementing the assessment protocol.

Acknowledgments

The thought-provoking comments of an anonymous reviewer greatly contributed to clarification of the ideas presented in this article. Portions of this work were performed under U. S. Environmental Protection Agency (EPA) contract 0V1036NTSX and Raytheon Corporation contract 99-9011-P1114C. This work has not been reviewed by EPA, Raytheon, or the U.S. Geological Survey and does not necessarily reflect the views of either agency or Raytheon.

References

Bellhouse, D.R., 1988. A brief history of sampling methods, *Handbook of Statistics, Vol. 6: Sampling* (P.R. Krishnaiah and C.R. Rao, editors), Elsevier Science Publishers, New York, N.Y., pp. 1–14.

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

Burnham, K.P., D.R. Anderson, and J.L. Laake, 1980. *Estimation of Density for Line Transect Sampling of Biological Populations*, Wildlife Monographs 72, The Wildlife Society, Washington, D.C., 202 p.

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

Cochran, W.G., 1977. *Sampling Techniques, Third Edition*, John Wiley & Sons, New York, N.Y., 428 p.

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

Congalton, R.G., and K. Green, 1999. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, Lewis Publishers, Boca Raton, Florida, 137 p.

Crist, P., and R. Deitner, 1998. *Assessing land cover map accuracy, Version 2.0.0, A Handbook for Conducting Gap Analysis*, USGS Gap Analysis Program, Moscow, Idaho, URL: http://www.gap.uidaho.edu/handbook/LandCoverAssessment/default.htm.

Czaplewski, R.L., 1999. Accuracy assessments and areal estimates using two-phase stratified random sampling, cluster plots, and the multivariate composite estimator, *Quantifying Spatial Uncertainty in Natural Resources: Theory and Applications for GIS and Remote Sensing* (H.T. Mowrer and R.G. Congalton, editors), Ann Arbor Press, Chelsea, Michigan, pp. 79–100.

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

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

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

Hansen, M.H., W.N. Hurwitz, and W.G. Madow, 1953. *Sample Survey Methods and Theory, Volumes I and II*, John Wiley & Sons, New York, N.Y., 970 p.

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

Horvitz, D.G., and D.J. Thompson, 1952. A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, 47:663–685.

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

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

Jensen, J.R., 1996. *Introductory Digital Imaging Processing: A Remote Sensing Perspective, Second Edition*, Prentice Hall, Upper Saddle River, New Jersey, 318 p.

Jessen, R.J., 1978. *Statistical Survey Techniques*, John Wiley & Sons, New York, N.Y., 520 p.

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

Kish, L., 1965. *Survey Sampling*, John Wiley & Sons, New York, N.Y., 643 p.

Murthy, M.N., 1967. *Sampling Theory and Methods*, Statistical Publishing Society, Calcutta, India, 706 p.

Paulsen, S.G., R.M. Hughes, and D.P. Larsen, 1998. Critical elements in describing and understanding our nation's aquatic resources, *Journal of the American Water Resources Association*, 34:995–1005.

Peterson, S.A., N.S. Urquhart, and E.B. Welch, 1999. Sample representativeness: A must for reliable regional lake condition estimates, *Environmental Science and Technology*, 33:1559–1565.

Raj, D., 1968. *Sampling Theory*, McGraw-Hill, New York, N.Y., 302 p.

Sarndal, C.E., B. Swensson, and J. Wretman, 1992. *Model-Assisted Survey Sampling*, Springer-Verlag, New York, N.Y., 694 p.

Schreuder, H.T., T.G. Gregoire, and G. Wood, 1993. *Sampling Methods for Multiresource Forest Inventory*, John Wiley & Sons, New York, N.Y., 446 p.

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.

———, 1996b. Estimating the kappa coefficient and its variance under stratified random sampling, *Photogrammetric Engineering & Remote Sensing*, 62:401–407.

———, 1996c. Cost-effective, practical sampling strategies for accuracy assessment of large-area thematic maps, *Spatial Accuracy Assessment in Natural Resources and Environmental Sciences: Second International Symposium*, 21–23 May, Fort Collins, Colorado, General Technical Report RM-GTR-277, USDA Forest Service, Fort Collins, Colorado, pp. 485–492.

———, 1997. Selecting and interpreting measures of thematic classification accuracy, *Remote Sensing of Environment*, 62:77–89.

———, 1999. Basic probability sampling designs for thematic map accuracy assessment, *International Journal of Remote Sensing*, 20:2347–2366.

———, 2000. Practical implications of design-based sampling inference for thematic map accuracy assessment, *Remote Sensing of Environment*, 72:35–45.

Stehman, S.V., and R.L. Czaplewski, 1997. Basic structures of a statistically rigorous thematic accuracy assessment, *Proceedings of the American Society for Photogrammetry & Remote Sensing*, 07–10 April, Seattle, Washington, 3:543–553.

———, 1998. Design and analysis for thematic map accuracy assessment: Fundamental principles, *Remote Sensing of Environment*, 64:331–344.

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

Stuart, A., 1984. *The Ideas of Sampling, Third Edition*, Oxford University Press, New York, N.Y., 91 p.

Thompson, S.K., 1990. Adaptive cluster sampling, *Journal of the American Statistical Association*, 85:1050–1059.

———, 1992. *Sampling*, John Wiley & Sons, New York, N.Y., 343 p.

Zhu, Z., L. Yang, S.V. Stehman, and R.L. Czaplewski, 2000. Accuracy assessment for the U. S. Geological Survey Regional Land-Cover Mapping Program: New York and New Jersey region, *Photogrammetric Engineering & Remote Sensing*, 66(12):1425–1435.

(Received 11 November 1999; accepted 17 May 2000; revised 14 September 2000)

SHAPE THE FUTURE
Careers in Imaging and Geospatial Information Science and Technology