Practical Implications of Design-Based Sampling Inference for Thematic Map Accuracy Assessment

*Stephen V. Stehman**

Sampling inference is the process of generalizing from sample data to make statements or draw conclusions about a population. Design-based inference is the inferential framework commonly invoked when sampling techniques are used in thematic map accuracy assessment. The conceptual basis of design-based inference is described, followed by discussion of practical implications of design-based inference, including (1) the population to which the inferences apply, (2) estimation formulas and their justification, (3) interpretation of accuracy measures, (4) representation of variability, (5) effect of spatial correlation, and (6) role of probability sampling. Design-based inference is contrasted with model-based inference, another inferential framework frequently invoked in statistics. ©*Elsevier Science Inc., 2000*

INTRODUCTION

Thematic map accuracy assessments are motivated by a variety of objectives. Describing classification error is often the primary objective, and reporting an error matrix (e.g., Story and Congalton, 1986) along with associated summary measures (e.g., overall proportion correctly classified, and user's and producer's accuracies) addresses this descriptive objective. Other accuracy assessment objectives include describing the spatial pattern of classification errors, assessing effects of environmental conditions on map accuracy, comparing different mapping techniques or classification schemes, or selecting the best land-cover map for a particular region. Accuracy assessments are based on comparing the map land-cover label to the reference (true) land-cover label at the same spatial location. Statistical sampling plays an important role because it is impractical to obtain the reference land cover for a census of the entire region. Although the inference issues apply to a map of either pixels or N polygons, to simplify the discussion, the presentation will focus on a land-cover map consisting of N pixels, with each pixel assigned to one land-cover class. A sample of n pixels is selected, and the reference land-cover class is obtained for each sampled pixel. Sampling inference is defined as the process of generalizing from this sample to make statements or draw conclusions about the population.

This article describes the practical implications of design-based inference, the common framework applied to sampling problems, for the design, analysis, and interpretation of accuracy assessment data. Design-based inference differs fundamentally from the model-based approach to inference commonly invoked when applying familiar statistical techniques such as analysis of variance, regression, and chi-square analysis of contingency tables. Contrasting design-based with model-based inference reveals that the assumptions and analyses of design-based inference are different from those of model-based inference, and applying model-based thinking to accuracy assessment problems can lead to some of the misconceptions and vague notions that occasionally hinder the practice of accuracy assessment. By understanding design-based inference, we are better able to recognize the assumptions and conditions critical to rigorous statistical inference for accuracy assessment applications. The goal is to adhere to those criteria required for rigorous inference, and to avoid those restrictions and assumptions that are unnecessary to valid sampling inference.

DESIGN-BASED INFERENCE

Design-based inference is the classical approach to sampling inference (Cochran, 1977; Kish, 1965; Murthy, 1967; Thompson, 1992; Yates, 1981) when the objective is to describe characteristics (e.g., means, proportions, totals) of a

* State University of New York, College of Environmental Science and Forestry, Syracuse, NY

Address correspondence to S.V. Stehman, State University of New York, College of Environmental Science and Forestry, 320 Bray Hall, Syracuse, NY 13210. E-mail: svstehma@mailbox.syr.edu

Received 22 December 1998; revised 19 July 1999.

REMOTE SENS. ENVIRON. 72:35–45 (2000)
©Elsevier Science Inc., 2000
655 Avenue of the Americas, New York, NY 10010

0034-4257/00/$–see front matter
PII S0034-4257(99)00090-5

real, explicitly defined population (Särndal et al., 1992, p. 514). Typical applications of design-based inference address sampling problems, where a sampling problem is defined as one in which (1) interest focuses on collective properties (parameters) of a real, existing population, (2) the collective properties do not need to be known with certainty, and (3) the attributes of all individual elements of the population are not required to be known. Translating this definition to the accuracy assessment setting, a sampling problem has as its motivating objective estimation of one or more collective properties (such as overall accuracy or producer's accuracies) of a target population of N pixels or polygons; these accuracies need not be known exactly (i.e., some uncertainty in the reported value is acceptable); and the accuracy of every single pixel or polygon displayed on the map is not required. Accuracy assessment of large-area, land-cover maps (Belward, 1996; Edwards et al., 1998; Muller et al., 1998; Zhu et al., 1999) is a typical example in which the conditions of a sampling problem are met.

In design-based inference, the population to which inference is intended to apply consists of the N pixels displayed on the map. A simple example of an attribute or response observed on each pixel is to assign $y_i=1$ to pixel i if it is correctly classified, and to assign $y_i=0$ if the pixel is incorrectly classified. In design-based inference, y_i is viewed as a fixed value, not a random variable. A parameter is a collective property of the population of N pixels (i.e., a parameter is a number describing a population). For example, the parameter representing the proportion of pixels correctly classified is $P=N_c/N$, where $N_c=\Sigma_{i=1}^{N}y_i$ is the number of pixels correctly classified. Similarly, user's accuracy is a proportion, the number of pixels mapped correctly as land-cover type A divided by the number of pixels mapped as A. Other parameters may be derived from the fuzzy classification approach to accuracy assessment. For example, the proportion of matches for each land-cover class, as defined by the RIGHT operator or the MAX operator developed from the linguistic scale described by Gopal and Woodcock (1994), represents collective properties of a population of N pixels. The arithmetic mean of the DIFFERENCE operator applied to all N pixels constitutes another collective property.

Variability in design-based inference arises from the randomization invoked in the sampling design. That is, although y_i is a fixed value, y_i is unknown unless pixel i is selected in the sample, so variability derives from the randomization associated with which pixels are selected in a particular realization of the sampling protocol. A practical consequence of this representation of variability is that the sampling design plays a pivotal role in design-based inference. For example, suppose the design is simple random sampling with sample size n and $\hat{P}$ is an estimate of the population proportion P. Many different simple random samples are possible, and the particular subset of elements selected in each sample will determine $\hat{P}$. The relevant uncertainty in design-based inference is the variability of $\hat{P}$ over the set of possible simple random samples: how repeatable is the estimate $\hat{P}$ if a different simple random sample is obtained? If stratified random sampling is used, variability of $\hat{P}$ over the possible stratified random samples differs from the variability of $\hat{P}$ over all possible simple random samples. The dependence of the variability of $\hat{P}$ on the design is a distinguishing characteristic of design-based inference. Stuart (1984) provides a lucid description of the concept of sampling variation in design-based inference.

MODEL-BASED INFERENCE

The term *model-based inference* derives from the starting point of such an inference, a model or process that generates an attribute or response for each of the N population pixels. In contrast to design-based inference, which applies to real, observable populations, model-based inference is often directed toward more hypothetical constructs or models that are a "product of human imagination" (Smith, 1997, p. 26). The model is viewed as generating "a long series of realizations" of populations of N pixels (Särndal et al., 1992, p. 534) potentially spanning a wide variety of settings, including future realizations of the process. The term *superpopulation* is sometimes used (Särndal et al., 1992, p. 22) to indicate that the inferential objective extends beyond the population of N elements of a single realization of the model. Because the characteristics of this superpopulation are usually specified by a model, the phrase *model-based inference* will still be applicable.

In accuracy assessment, model-based inference may be invoked to infer general properties of a particular classification method (e.g., neural network, linear discriminant analysis, or k–nearest neighbor). The objective is to characterize the performance of the classifier in general, not just performance for a single application or realization. The inference is desired to generalize to other regions, to the same region at a different time, or to a different land-cover classification scheme. Comparing different classification techniques is another prototypical application of model-based inference. To conclude that one mapping technique is better than another is an inference about the two classification techniques in general, not just one realization of each technique. The conclusion drawn from the inference is intended to apply to a relatively broad set of conditions and future applications of the two techniques.

In model-based inference, the classification error attributes of the pixels are represented as random variables generated by a statistical model. For example, the model may specify that the classification generates pixels such that the probability of a correct classification is p, and that the classification error process operates independently on each pixel (p will be used to denote a model parameter, and P will denote a parameter of a particular population of N pixels). Under this model, y_i is a Bernoulli random variable (i.e., $y_i=1$ if the pixel is classified correctly, $y_i=0$ otherwise). Although y_i has the same numerical form here as in the design-based mode, in model-based inference y_i is regarded as a random variable, not a fixed value. If a sample of n pixels is generated by this model, the random variable $X=$(number of correct classifications) has a binomial distribution with parameters n and p. The interpretation of the parameter p and the representation of uncertainty are both model dependent. For the binomial random variable, for example, the variance of X is $V(X)=np(1-p)$, and if $\hat{p}=X/n$ is an estimator of the model parameter p, the variance of $\hat{p}$ is $V(\hat{p})=p(1-p)/n$. If the model is not correct—for example, if classification errors tend to cluster rather than to be independent—then these binomial variance formulas no longer apply, because the independence assumption of the binomial model is violated.

COMPARING THE TWO INFERENCE FRAMEWORKS

In design-based inference, the population of N pixels is a well-defined, tangible entity. If a census of the true land cover is obtained, then y_i is known for all N pixels, and the proportion of pixels correctly classified, $P=\Sigma_{i=1}^{N}y_i/N$, as well as other accuracy parameters, would be known exactly. Of course, it is impractical to obtain a census, but the point is that parameters defined in the design-based framework are quantities representing a real population for which each element of this population is potentially observable. In contrast, the populations to which model-based inferences apply are often less tangible. Even if we had a census of reference data for a particular map, if the objective is to infer accuracy of the process generating land-cover classification error, uncertainty would still remain, because a model parameter "is part of a hypothetical construct, therefore it can never be calculated exactly" (Särndal et al., 1992, p. 515). For example, suppose the objective of inference is to determine the accuracy of land-cover classifications resulting from a neural network classifier. The N pixels of a single land-cover map represent only a portion of the potential outputs of this classification process. Even provided with a census of reference classifications for these N pixels, uncertainty about the accuracy of the process would remain, because the full possible range of realizations of this neural network classifier has not been observed. The superpopulation or process to which model-based inferences apply generally does not consist of elements that are all observable.

A useful guideline for determining whether to operate in the design-based or model-based framework is to ask the question: "If a census of ground truth or reference data were available, would the question posed by the objective be answered with certainty?" If yes, the inference objective pertains to a real population of N pixels as it exists "right now", and design-based inference is applicable (Särndal et al., 1992, p. 515). If the question is answered no, the inferential objective is likely a process or more expansive superpopulation generalizing beyond the N pixels of the observable population, and model-based inference applies. A few examples illustrate these ideas.

Suppose a wildlife scientist is primarily interested in one cover type (e.g., oak woodland) because this type is important habitat for the species under investigation. If the accuracy assessment objective is to determine user's accuracy of oak woodland, then a complete census of reference data would reveal the answer with certainty. For example, if 5,068 pixels are labeled as oak woodland on the map, and 4,288 are in fact oak woodland on the ground, then 4,288/5,068=0.846 is the true user's accuracy. Provided with a census, the scientist knows user's accuracy of oak woodland for this one map with certainty. In practice, the user's accuracy estimate would be derived by sampling from the 5,068 available oak-woodland pixels, and design-based inference would be used to generalize the results from this reference sample to the entire population. Because the mapping application depends on this particular map to determine potential habitat, the scientist's objective is served by evaluating the accuracy of this map. The objective is not to learn how well the same mapping process would work when applied to another region or at a future point in time.

As a second example, suppose the objective is to compare two mapping techniques, one a supervised classification, the other an unsupervised classification combined with a post-classification sorting of the unsupervised classes. A single region is mapped using both techniques, and the accuracy of the two resulting maps is compared. If a census of reference data were obtained, we would have the exact accuracy measures for a single realization of each mapping technique. The census information allows us to answer with certainty the question of which technique is better for this region. Consequently, deciding which of two techniques (or two maps) is better for classifying a particular region is addressed within the design-based framework. However, provided with a census of reference data for just this single region, we would not have a certain answer to the question of which mapping technique is better in general. This broader comparison objective is addressed inadequately with data from only a single case study. When inferring the comparative merits of the two mapping techniques in general, uncertainty would remain about which technique is better, because a particular region represents only one realization of the settings in which these techniques potentially could be applied. Our ability to infer or generalize from a single region is limited, suggesting that a general comparison of two mapping techniques requires more than one application of each technique.

As a final example, suppose the objective is to evaluate sources of classification error. Are classification errors attributable to elevation, aspect, hydrological regime, image date, or the individual carrying out the classification? Is uncertainty primarily associated with the training procedure and generating boundary pixels (Wang and Howarth, 1993)? If the objective is to quantify sources of classification error for a particular map, then design-based inference applies and a sample from a single map provides the necessary data for the inference. Conversely, if the goal is to identify sources of error for the classification process in general, sampling from more than a single map is necessary to more broadly represent the process.

Practical realities of accuracy assessment may require incorporating elements of both inferential frameworks in the analysis. For example, when missing data (e.g., nonresponse due to denied access to sample locations) or measurement error in the reference data (e.g., photointerpreter error) exist, features of model-based inference may need to be incorporated into the analysis even if design-based inference is the desired framework for the application. See Särndal et al., 1992, Chapters 15 and 16, for an overview of these techniques in a general sampling context. The extensive general discussions of design- and model-based inference provided by De Gruijter and Ter Braak (1990), Gregoire (1998), Hansen et al. (1983), Overton (1993), and Smith (1997) provide further explanations of the distinctions and similarities of the two modes of inference.

INTERPRETING ACCURACY MEASURES

Accuracy parameters defined in the design-based framework are typically interpreted as collective properties characterizing the N pixels comprising the population. Alternatively, these parameters may be interpreted as probabilities. For example, the overall proportion correctly classified, P, is interpretable as the probability that a randomly selected pixel is correctly classified. User's accuracy is interpretable as a conditional probability: given that a randomly selected pixel is classified as category A by the map, what is the probability that the pixel is actually category A on the ground? The feature common to these probabilistic interpretations is that the probability applies to the process of *selecting a pixel at random.* Parameters such as P, and user's and producer's accuracies, are not interpretable as the probability that a pixel at a specific (i.e., not randomly selected) location is classified correctly. Because in design-based inference y_i is a fixed value, not a random variable, the probability interpretation applies to the pixel selection process, not to the observation y_i. Although many design-based parameters are proportions, these proportions cannot be translated into probability statements about the accuracy of a pixel at a predetermined location. Smith (1997, pp. 35–36) provides additional discussion on these matters of interpretation in design-based inference.

Defining location-specific probabilities focuses on model-based features characterizing the process generating classification error. Such probabilities may be derived from the classification procedure. For example, maximum likelihood techniques generate class membership probabilities of the pixel belonging to each of the possible land-cover classes. These probabilities may be used as a measure of uncertainty of the land-cover class assigned, and therefore represent a characteristic of the classification process. However, because these class membership probabilities are not the result of comparing the map label to a reference label, they are not interpretable as accuracy measures (Canters, 1997; Zhu, 1997). Corves and Place (1994, p. 1284) define these probabilities associated with the classification process as measures of reliability. Because reliability is known for all N pixels, no sampling inference is required to obtain a summary measure of reliability for the population.

Some accuracy measures are based on models, and consequently their interpretation is model dependent. The most commonly used model-dependent parameters, such as κ (Congalton et al., 1983; Rosenfield and Fitzpatrick-Lins, 1986) and τ (Ma and Redmond, 1995), invoke a model to define "random chance agreement." Different models lead to different measures. For example, κ assumes independence of the map and reference classifications, and assumes that the mapped land-cover proportions have been constrained to produce the observed map proportions in each class. Parameters defined on the basis of model structures can still be estimated in the design-based framework and interpreted as collective properties of a population of N pixels. Stehman (1996a) illustrates this approach in deriving an estimator for κ and its standard error under stratified random sampling. The consistency criterion described in the following section applies to estimation in the design-based framework even when the definition of the parameter is model dependent.

The use of model-dependent accuracy measures poses some interesting inference questions. Turk (1979) states that a chance-agreement adjustment is not necessary for describing the accuracy of a particular map (a typical design-based objective), because such correct classifications are a windfall gain to this map and should not be subtracted from the map's reported accuracy. To describe accuracy of an existing map, factoring in how or why the classifier correctly labeled a particular pixel is not the objective and a model-dependent, chance-agreement adjustment for accuracy is not relevant. Conversely, if the objective is to predict how accurately this classifier will perform in future applications (a typical objective in model-based inference), then a correction for chance agreement is better motivated. Prediction to future applications of the classifier focuses on the classification process and model structures representing chance agreement may be relevant, because future success of the classifier depends on how or why the classifier performs well. Therefore, a measure corrected for chance agreement may better reflect the true future performance of the classifier. The question of how best to model chance agreement in the classification process remains unresolved.

ESTIMATION

In design-based inference, estimation is founded on the consistency criterion. By definition, an estimator is consistent if it equals the population parameter when the sample size, n, equals the population size, N (Cochran, 1977, p. 21). A practical implication of adhering to the consistency criterion is that the sampling design is relevant to estimation: different sampling designs may require different formulas to estimate consistently the same parameter. For example, if the design is stratified random sampling, applying formulas appropriate for simple random sampling results in misleading estimates because the consistency criterion has been violated (Stehman, 1996b). Variance estimation formulas should also satisfy the consistency criterion, and consequently these formulas also are specific to the sampling design used. Stehman (1995) and Stehman and Czaplewski (1998) provide additional discussion of consistent design-based estimation for accuracy assessment, and Overton (1993) and Särndal et al. (1992, sec. 5.3) supply general overviews of this topic.

Gregoire (1998) notes that model-based inferences use estimation methods such as maximum likelihood, least squares, and minimax risk. Consequently, estimator formulas in model-based inference are derived using different criteria from those of design-based inference. In some situations, the estimation formulas derived from the two frameworks are similar. For example, suppose the sampling design is simple random sampling and the model-based approach uses the binomial model. The resulting estimation formulas are the same as those derived in the design-based mode, except when the sampling fraction, $f=n/N$, is large, in which case the design-based variance estimates incorporate a finite population correction term, $1-n/N$, absent from the model-based variance estimates. Stratified random sampling also leads to similar estimation formulas. In the model-based framework, Card (1982) used a multinomial model for the cell proportions within each stratum to derive maximum likelihood estimators for several accuracy parameters. Green et al. (1993) derived an estimator of "producer's risk", also using the multinomial model, and applying Bayes's theorem. More complex models could be constructed to derive estimators that account for the extra variation contributed by clustering.

The assumptions required for estimation differ for design-based and model-based inferences. The consistency criterion of design-based inference does not require that the data follow a specified probability distribution, nor does it require assuming that the observations are independent. Recall that y_i is not a random variable in the design-based framework, so y_i has no probability distribution associated with it. Further, the usual random variable definition of independence does not apply to the observations y_i and y_j in the design-based framework because y_i and y_j are fixed values, not random variables. Model-based inference typically assumes a probability distribution for y_i as part of the model structure. This probability distribution is usually chosen to reflect features of the sampling design, such as strata or clusters. However, the modeler may choose to ignore the sampling design entirely. For example, if the model assumes independent observations and this specification is valid, then the usual concern with correlated observations within clusters in a cluster sampling design is not warranted. In reality, spatial correlation will undoubtedly be present and the independence assumption required in the model-based framework will be violated for a cluster sampling design. In model-based inference, accuracy assessment data are not inherently distributed as binomial, multinomial, or normal, but rather the data are modeled as such. Model validity should be evaluated as part of the data analysis protocol to determine if the specified model is appropriate for the application.

CONFIDENCE INTERVALS

In the design-based framework, confidence intervals are usually constructed via the general form $\hat{\theta}\pm z_a\mathrm{SE}(\hat{\theta})$, where $\hat{\theta}$ is an estimated accuracy parameter, z_a is a percentile from the standard normal distribution, and $\mathrm{SE}(\hat{\theta})$ is the estimated standard error of $\hat{\theta}$ (Särndal et al., 1992, sec. 2.11). The sample size is assumed to be sufficiently large that the sampling distribution of the estimator, $\hat{\theta}$, is approximately normal, as justified by the central limit theorem. The assumption required for constructing confidence intervals is not that accuracy data are normally distributed or continuous, but rather that the sampling distribution of the estimator is approximately normal. For example, to construct a confidence interval for the parameter P, the assumption is not that y_i is normally distributed, but that the estimator $\hat{P}$ is approximately normally distributed. If the sample size n is large, it is the approximate normality of $\hat{P}$ that is assured by the central limit theorem. What constitutes a sample large enough to invoke the central limit theorem depends on the estimator, the particular population being sampled, and the sampling design. It is difficult to provide practical guidelines on when n is sufficiently large. Another key element to the validity of the confidence interval technique is applying the consistency criterion to estimating standard errors. The general confidence interval formula, $\hat{\theta} \pm z_{\alpha}\mathrm{SE}(\hat{\theta})$, still applies regardless of the sampling design, but the specific formula for $\mathrm{SE}(\hat{\theta})$ will change depending on whether the design is simple random sampling, stratified random sampling, or cluster sampling.

Model-based procedures for constructing confidence intervals exist. For example, if the sampling design is simple random sampling, confidence intervals may be constructed from the binomial distribution (Thomas and Allcock, 1984; Morisette and Khorram, 1998) if the sampling fraction $f=n/N$ is small. The model-based nature of these intervals should be recognized, because the procedures may not be robust to deviations from the specified model. For example, if the design is cluster sampling rather than simple random sampling, the independence assumption of the binomial model is likely violated, and confidence intervals constructed from the binomial distribution will not have the specified nominal coverage. The binomial model is also inappropriate for stratified sampling with equal allocation and small sample sizes in each stratum when constructing confidence intervals for parameters of the full population (i.e., combining strata).

SAMPLING INFERENCE AND SPATIAL ISSUES

Spatial Variation in Classification Error

The descriptive objectives of accuracy assessment may include estimating subregional accuracy, where the subregions are ecoregions, states, or arbitrarily defined areas within the map. Accuracy may vary among these subregions, and estimates summarizing accuracy of the full region will not capture this subregional spatial variation. Design-based inference applies to subregional estimates, because any subregion represents a subpopulation of N' pixels ($N'<N$), and the usual accuracy measures can be defined as collective properties representing the subregion. The sample data within each subregion are then used to estimate accuracy, and these subregional estimates provide a description of the spatial variation in accuracy. A sufficient number of sample pixels must be contained within the subregion to provide reasonably precise estimates, so for small areas, estimator precision is a concern. For example, suppose a land manager is interested in the map's accuracy for a 25-hectare management unit. Only a few sample points from a large-area accuracy assessment will likely be contained in the 25-hectare area, so the estimated accuracy will be imprecise.

Design-based inference provides no mechanism for predicting the accuracy of unsampled pixels. Prediction requires a model to relate the data from the sampled pixels to those pixels not sampled. Consequently, a model-based approach may be required to provide accuracy estimates for small subregions in which few or no sampled pixels are found, or to predict accuracy for individual pixels. The topic "small area estimation" (cf. Särndal et al., 1992, sec. 10.8) has potential application when accuracy of a small subregion is desired. The prediction problem could be addressed by constructing a model for predicting accuracy as a function of elevation, aspect, and other auxiliary variables affecting classification error.

Describing Spatial Patterns of Classification Error

Mapping or visualizing classification error is an objective for which a model-based approach is potentially more informative than a design-based approach. A simple representation of the spatial pattern of classification error is available by displaying the spatial locations of the correct and incorrect classifications for the reference sample. This display also provides information on the spatial distribution of the sample itself. No formal inference is implied by such a graphical depiction of the data. Often a more general spatial representation of classification error will be desired, and model-based approaches to predict accuracy at unsampled locations become necessary. A simple method is to assign to an unsampled pixel the value of y (i.e., $y_i=1$ or $y_i=0$) of the spatially closest sample pixel. The model supporting this method assumes that pixels in close proximity to one another are more similar to each other than they are to pixels more distant.

Kriging (i.e., spatial interpolation) is a more sophisticated modeling approach for predicting y at unsampled locations. The interpolated values can be combined with the sample data to construct a contour plot of y. Indicator kriging (Isaaks and Srivastava, 1989, Chapter 18) is one option applicable for the 0–1 representation of classification error. Steele et al. (1998) report this method results in a difficult to interpret, highly variable surface when the sample data are sparse, and instead recommend a bootstrap approach to derive misclassification probabilities prior to applying kriging. A misclassification probability (taking on a value between 0 and 1) is computed for each sampled pixel, and kriging is used to interpolate these probabilities for unsampled pixels. The Steele et al. (1998) contour plots display a smooth, visually appealing surface for the spatial pattern of classification error.

The misclassification probabilities derived at each sample point from this bootstrap method illustrate the contrast between characterizing the classification process and describing accuracy of a realized map. Consider a particular sample pixel. In the design-based framework, a probability is not attached to the accuracy of this pixel, and because the pixel is in the reference sample, it is known whether the pixel is correctly classified. Conversely, the bootstrap approach generates a probability associated with the classification process for this sample pixel: if the process were repeated multiple times, what percent of the time would this pixel be classified correctly? This establishes the model-based context of the probabilities obtained via the bootstrap approach. The bootstrap misclassification probabilities are analogous to the Corves and Place (1994) reliability measures. An appealing feature of the model used by the bootstrap is that it requires no explicit specification of a probability distribution. Instead, this probability distribution is established empirically by the sample data on which the bootstrap resampling is performed.

Fisher (1997) describes another approach to representing spatial variation in classification error. Spatial variation can be visualized either by randomizing the land-cover type displayed at a particular location, or by randomizing the pixel locations at which the land-cover type is changed. The visualization technique permits the option of using standard accuracy measures such as overall accuracy or user's accuracies to guide the randomization. In this application, the randomization structure functions as the model used to create information at the unsampled pixel locations.

Spatial Correlation

Spatial correlation, defined as the degree to which classification errors cluster spatially, may be attributable to spatial patterns in land cover or may result from features of the classification process such as edge matching, different imagery dates, or different technicians classifying different regions of the map. Spatial correlation does not change the analysis techniques of design-based inference. The estimation and standard error formulas remain the same no matter the magnitude of the spatial correlation, but the actual values of the estimates, particularly standard errors, are affected by spatial correlation. The presence of spatial correlation does not violate any assumptions in the design-based framework because the derivation of the estimators and variances does not assume independence of the observations (y_i).

Taking spatial correlation into account is useful when planning the sampling design because spatial correlation influences precision, particularly of systematic sampling, cluster sampling, and stratified random sampling (when the strata are geographic regions). A strong, positive spatial correlation produces a high intracluster correlation, resulting in poor precision for cluster sampling. Positive spatial correlation means that neighboring pixels have similar values of y, so pixels within each cluster are likely to carry redundant information. The effect of high intracluster correlation on precision is diminished if the cluster size is small, but small clusters also negate the main advantage of cluster sampling: reduced travel costs due to the close proximity of the sampling units within a cluster. Positive spatial correlation of classification error generally works to the advantage of systematic sampling. This design has low variance if neighboring elements in the population tend to have similar y values (Cochran, 1977, p. 208). By definition, a positive spatial correlation provides such a favorable pattern. Periodic spatial patterns in classification error may sometimes arise, resulting in a negative spatial correlation at particular distances. If the systematic sampling interval selected matches this distance (i.e., the systematic sampling interval is in phase with the periodicity), the systematic design will have poor precision. The more that is known about the spatial pattern of classification error, the less likely this problem is to occur. When geographic strata are used, a positive spatial correlation is again beneficial. Stratification enhances precision when within-stratum variation is low. Therefore, positive spatial correlation indicates that geographically nearby elements have similar response values, thus producing the favorable within-stratum homogeneity.

Congalton (1988a), Lo and Watson (1998), and Pugh and Congalton (1997) provide case studies in which spatial correlation of classification error has been quantified over a large area. How to best characterize the effect of spatial correlation depends on the sampling design. A spatial correlation measure such as the join count statistic does not directly provide design-specific information. Instead, the effect of spatial correlation is better captured by the intracluster correlation for cluster sampling (Cochran, 1977, p. 209) and the within-stratum variance for stratified random sampling (Cochran, 1977, p. 90). In practice, the actual spatial correlation will be unknown. Sample design planning may proceed based on a crude estimate of spatial correlation obtained from a small pilot study, or more typically based on a guess of the magnitude of spatial correlation.

Spatial correlation affects model-based inference because it represents a departure from the independence assumption usually invoked in model-based analyses (e.g., Cressie, 1991, Secs. 4.3.2, 5.7; Fingleton, 1983; Griffith, 1978). Even the simpler statistical models applied to accuracy assessment (e.g., the binomial and multinomial models) assume independence. For model-based analyses, spatial correlation affects both the precision of estimators and the estimation of that precision. Typically, the effect of positive spatial correlation is to increase the variance of an estimator relative to when the observations are independent and to produce underestimates of variance. Underestimating variance results in confidence intervals for which the true coverage is lower than the stated nominal coverage (e.g., a stated 95% confidence interval has actual coverage of, say, 82%) and in hypothesis tests that have actual Type I error higher than the stated nominal level (α).

Several authors have expressed concern about the effect of spatial correlation on sampling design and analysis in accuracy assessment: (1) "Each error matrix built for accuracy analysis must satisfy the following assumptions: Pixels are sampled independently . . ." (Sharma and

Sarkar, 1998, p. 276); (2) "This [using 3×3 pixel blocks] introduced the possibility of spatial autocorrelation between samples in these categories which is a potential source of bias" (Muller et al., 1998, p. 623); (3) "Another problem related to cluster sampling lies in the spatial autocorrelation of neighboring pixels which tends to bias sample estimates with increasing cluster size" (Corves and Place, 1994, p. 1285); (4) "Spatial autocorrelation may be a source of bias in error matrix construction" (Hess and Bay, 1997, p. 316); and (5) "The sample procedure utilized should minimize the effects of spatial autocorrelation . . ." (Dicks and Lo, 1990, p. 1248). The validity of these statements is dependent on the inference framework. Because design-based inference does not assume independent observations, spatial correlation will not bias accuracy estimates, and the sampling design need not be chosen to avoid spatial correlation. The sampling design may be selected to mitigate the effect of spatial correlation on precision (e.g., by choosing small cluster sizes or using two-stage cluster sampling) or even to take advantage of spatial correlation to enhance precision (e.g., systematic sampling with a well-chosen sampling interval). When operating in the design-based framework, we need to be careful not to impose unnecessary restrictions on the sampling design or assumptions on the analysis by misplaced concern with the effect of spatial correlation. Assertions of bias and other flaws attributed to lack of independence arising from spatial correlation are irrelevant in design-based inference, but Gregoire (1998, Remark 3) notes that such assertions are still common. The potential negative impacts of spatial correlation on model-based inferences not incorporating this spatial structure are legitimate concerns.

To summarize, in the design-based framework, spatial correlation does not require changing the formulas used to estimate accuracy parameters and their standard errors, and the consistency criterion remains the guiding principle for estimation. Considering the spatial pattern of classification error is relevant when choosing a sampling design. Positive spatial correlation is generally favorable to stratified sampling (with geographic strata) and systematic sampling, and unfavorable to cluster sampling. Congalton (1988b) reviews the effect of spatial correlation on cluster sampling, and Moisen et al. (1994) provide a thorough description and analysis of the effect of spatial correlation on both cluster and systematic sampling. In the model-based framework, the presence of spatial correlation violates the independence assumption invoked by many analyses. The field of spatial statistics offers alternative models for analysis of spatially correlated data. De Gruijter and Ter Braak's (1990) discussion of model-based and design-based inference focuses specifically on these spatial correlation issues.

ROLE OF PROBABILITY SAMPLING

Statistical inference is predicated on the assumption that the sample is representative of the target population to which the inferences are intended to apply. In design-based inference, probability sampling provides the foundation justifying sample representativeness. Kish (1987, pp. 22–23) states that "probability sampling denotes the only feasible method recognized by survey samplers in most practical situations" to achieve a representative sample, where *representative* is defined as a mirror or miniature of the population. Because the population to which model-based inference applies is often hypothetical, it is not always possible, even when desirable, to obtain a probability sample from the target population. Consequently, the representativeness justification for model-based inference originates from the validity of the specified model.

Probability sampling is a critical element of design-based inference and the consistency criterion captures the essential probability sampling structure by requiring that the sampling design be accounted for by the estimators. Probability sampling may be relegated to a less prominent role in model-based inference. In the most extreme form of model dependence, model-based inference may be conducted ignoring the sampling design, with the inferences relying completely on the validity of the model (Särndal et al., 1992, p. 516): if the model is valid, it does not matter how the sample was selected because the model applies no matter which subset of elements is observed. Probability sampling is sometimes used for model-based inference to assure objectivity in sample selection, and to insure against the possibility of an incorrectly specified model. Balanced sampling has been proposed to protect inferences from model misspecification. This nonprobability sampling technique selects sample elements so that the sample mean of each of one or more control variables is approximately the same as the corresponding population mean for that control variable (Royall and Eberhardt, 1975; Särndal et al., 1992, pp. 517, 531, 536).

Inference from a nonprobability sampling design requires appealing to an assumption that is tantamount to a model. For example, using purposely selected training sites for accuracy assessment requires the assumption that these training sites are representative of the accuracy of the full population. Consequently, the statistical basis of inference ensured by the protocol of probability sampling is replaced by an assumption of representativeness. Because training sites are often selected from areas of homogeneous land cover, their accuracy is typically not representative of the accuracy of the population. Even if some of the training sites are withheld from the process used to develop the classification and then subsequently used in accuracy assessment, the problem still remains (Hammond and Verbyla, 1996). Inferences from samples selected from easily accessed or convenient locations similarly require a model assumption that these sites are representative of the target population.

Table 1. General Characteristics of Design-Based and Model-Based Inference

Characteristic	*Design-Based Inference*	*Model-Based Inference*
Population to which inference applies	N pixels (or polygons) of an existing map	Process or model generating classification errors
Probability sampling	Critical	Unnecessary if inferences completely dependent on model; recommended to ensure objectivity
Estimation	Consistent	Maximum likelihood, least squares, minimax risk, bootstrap
Variation		
Response, y	Fixed value (constant)	Random variable
Source of variability	Randomization component of sampling design	Model of random variable, y
Uncertainty present with census	No	Yes
Assumptions		
Probability distribution of the observations	No	Yes
y_i's independent	No	Yes
Spatial correlation	No effect on conduct of analysis; influences precision so relevant to planning	Violates independence assumption and may require incorporating spatial correlation in the model
Approximate normality of estimator	Yes, justified by large sample size	Yes, justified by large sample size
Interpretation	Collective properties of real population	Characteristics of model or process
Parameters		
Spatial variation of accuracy	Subregional estimates	Subregional estimates or mapping of classification error
Location (e.g., pixel) specific accuracy	No	Yes, via model

SUMMARY

Design-based inference has important practical implications on how accuracy assessments are designed, analyzed, and interpreted. Design-based inference focuses on a real, identifiable population of N pixels or polygons, such as that represented by a particular land-cover map at a given moment in time. Probability sampling and consistent estimation are the foundation of rigorous design-based inference, allowing generalization from the sample to the target population. In design-based inference, the observed response, y_i, is a fixed value, not a random variable, and variation is attributed to the randomization incorporated in the sampling design. No probability distribution is assumed for y_i and, because the sample observations are not assumed to be independent, design-based inference does not require special techniques when spatial correlation is present.

Model-based inference is often invoked when inference is intended to apply to a process or superpopulation more expansive than that represented by the N pixels or polygons of a single map. Comparing two classifiers in general is a prototypical question addressed by model-based inference. The question is rarely resolved by a single application of each classifier to one region. A probability distribution for y_i is often specified as part of the model, and because many model-based analyses assume that observations are independent, spatial correlation is an important consideration. Model-based techniques are often necessary when the analysis incorporates spatial registration error or measurement error in y (i.e., error in the reference land-cover classification). A model-based approach is also advantageous for constructing a map of classification error or predicting accuracy for individual map elements. General features of design-based and model-based inference are summarized in Table 1.

Both design-based and model-based inference have important roles in accuracy assessment, and both may be used within the same project. The inference framework selected for a particular objective will depend on the population to which inference is desired, and the analyst's choice of how to represent variation. Design-based inference has played the more prominent role in accuracy assessment because of the focus on descriptive objectives. Therefore, it is important for practitioners to be aware that the assumptions, estimation criteria, and interpretation of parameters in design-based inference differ from those of model-based inference. Applying model-based concepts to design-based inference may unnecessarily restrict the sampling design and analysis options, and can also lead to incorrect interpretation of accuracy assessment data. The key practical implications that arise when taking a design-based approach to inference are (1) the importance of satisfying the probability sampling design criterion, (2) the consistency criterion as the basis of estimation, (3) the absence of distributional (except for confidence interval construction) and independence assumptions, and (4) the interpretation of accuracy parameters as collective properties of the population.

Timothy Gregoire's review of an earlier version of this manuscript led to clarification of several concepts. Brian Steele and Gretchen Moisen provided many useful comments that significantly improved the presentation and focus of the article.

REFERENCES

Belward, A. S. (ed.) (1996), The IGBP-DIS Global 1km Land Cover Data Set: Proposal and Implementation Plans. IGBP-DIS Working Paper 13, Joint Research Centre, Space Applications Institute, Ispra, Italy.

Card, D. H. (1982), Using known map category marginal frequencies to improve estimates of thematic map accuracy. *Photogramm. Eng. Remote Sens.* 48:431–439.

Canters, F. (1997), Evaluating the uncertainty of area estimates derived from fuzzy land-cover classification. *Photogramm. Eng. Remote Sens.* 63:403–414.

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

Congalton, R. G. (1988a), Using spatial autocorrelation analysis to explore the errors in maps generated from remotely sensed data. *Photogramm. Eng. Remote Sens.* 54:587–592.

Congalton, R. G. (1988b), A comparison of sampling schemes used in generating error matrices for assessing the accuracy of maps generated from remotely sensed data. *Photogramm. Eng. Remote Sens.* 54:593–600.

Congalton, R. G., Oderwald, R. G., and Mead, R. A. (1983), Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques. *Photogramm. Eng. Remote Sens.* 49:1671–1678.

Corves, C., and Place, C. J. (1994), Mapping the reliability of satellite-derived landcover maps an example from the Central Brazilian Amazon Basin. *Int. J. Remote Sens.* 15(6): 1283–1294.

Cressie, N. A. C. (1991), *Statistics for Spatial Data*, Wiley, New York.

De Gruijter, J. J., and Ter Braak, C. J. F. (1990), Model-free estimation from spatial samples: a reappraisal of classical sampling theory. *Math. Geol.* 22:407–415.

Dicks, S. E., and Lo, T. H. C. (1990), Evaluation of thematic map accuracy in a land-use and land-cover mapping program. *Photogramm. Eng. Remote Sens.* 56:1247–1252.

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

Fingleton, B. (1983), Independence, stationarity, categorical spatial data and the chi-squared test. *Environ. Plann. A* 15: 483–499.

Fisher, P. (1997), The pixel: a snare and a delusion. *Int. J. Remote Sens.* 18:679–685.

Gopal, S., and Woodcock, C. (1994), Theory and methods for accuracy assessment of thematic maps using fuzzy sets. *Photogramm. Eng. Remote Sens.* 60:181–188.

Green, E. J., Strawderman, W. E., and Airola, T. M. (1993), Assessing classification probabilities for thematic maps. *Photogramm. Eng. Remote Sens.* 59:635–639.

Gregoire, T. G. (1998), Design-based and model-based inference in survey sampling: appreciating the difference. *Can. J. Forest Res.* 28:1429–1447.

Griffith, D. A. (1978), A spatially adjusted ANOVA model. *Geogr. Anal.* 10:296–301.

Hammond, T. O., and Verbyla, D. L. (1996), Optimistic bias in classification accuracy assessment. *Int. J. Remote Sens.* 17(6):1261–1266.

Hansen, M. H., Madow, W. G., and Tepping, B. J. (1983), An evaluation of model dependent and probability sampling inferences in sample surveys. *J. Amer. Stat. Assoc.* 78: 776–807.

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

Isaaks, E. H., and Srivastava, R. M. (1989), *An Introduction to Applied Geostatistics*, Oxford University Press, Oxford.

Kish, L. (1965), *Survey Sampling*, Wiley, New York.

Kish, L. (1987), *Statistical Design for Research*, Wiley, New York.

Lo, C. P., and Watson, L. J. (1998), The influence of geographic sampling methods on vegetation map accuracy evaluation in a swampy environment. *Photogramm. Eng. Remote Sens.* 64:1189–1200.

Ma, Z., and Redmond, R. L. (1995), Tau coefficients for accuracy assessment of classification of remote sensing data. *Photogramm. Eng. Remote Sens.* 61:435–439.

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* (W. K. Michener, J. W. Brunt, and S. G. Stafford, Eds.), Taylor & Francis, New York, pp. 159–176.

Morisette, J. T., and Khorram, S. (1998), Exact binomial confidence interval for proportions. *Photogramm. Eng. Remote Sens.* 64:281–283.

Muller, S. V., Walker, D. A., Nelson, F. E., Auerbach, N. A., Bockheim, J. G., Guyer, S., and Sherba, D. (1998), Accuracy assessment of a land-cover map of the Kuparuk River basin, Alaska: considerations for remote regions. *Photogramm. Eng. Remote Sens.* 64:619–628.

Murthy, M. N. (1967), *Sampling Theory and Methods*, Statistical Publishing Society, Calcutta.

Overton, W. S. (1993), Probability sampling and population inference in monitoring programs. In *Environmental Modeling with GIS* (M. F. Goodchild, B. O. Parks, and L. T. Steyaert, Eds.), Oxford University Press, New York, pp. 470–480.

Pugh, S. A., and Congalton, R. G. (1997), Applying spatial autocorrelation analysis to evaluate error in New England forest cover type maps derived from thematic mapper data. In *Proceedings of the 1997 ASCM/ASPRS Annual Convention*, ASPRS Technical Papers, Vol. 3, pp. 648–657.

Rosenfield, G. H., and Fitzpatrick-Lins, K. (1986), A coefficient of agreement as a measure of thematic classification accuracy. *Photogramm. Eng. Remote Sens.* 52:223–227.

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

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

Sharma, K. M. S., and Sarkar, A. (1998), A modified contextual classification technique for remote sensing data. *Photogramm. Eng. Remote Sens.* 64:273–280.

Smith, T. M. F. (1997), Social surveys and social science. *Can. J. Stat.* 25:23–44.

Steele, B. M., Winne, J. C., and Redmond, R. L. (1998), Estimation and mapping of misclassification probabilities for thematic land cover maps. *Remote Sens. Environ.* 66: 192–202.

Stehman, S. V. (1995), Thematic map accuracy assessment from the perspective of finite population sampling. *Int. J. Remote Sens.* 16:589–593.

Stehman, S. V. (1996a), Estimating the kappa coefficient and its variance under stratified random sampling. *Photogramm. Eng. Remote Sens.* 62:401–407.

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

Stehman, S. V., and Czaplewski, R. L. (1998), Design and analysis for thematic map accuracy assessment: fundamental principles. *Remote Sens. Environ.* 64:331–344.

Story, M., and Congalton, R. G. (1986), Accuracy assessment: a user's perspective. *Photogramm. Eng. Remote Sens.* 52: 397–399.

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

Thomas, I. L., and Allcock, G. McK. (1984), Determining the confidence level for a classification. *Photogramm. Eng. Remote Sens.* 50:1491–1496.

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

Turk, G. T. (1979), GT index: a measure of the success of prediction. *Remote Sens. Environ.* 8:65–75.

Wang, M., and Howarth, P. J. (1993), Modeling errors in remote sensing image classification. *Remote Sens. Environ.* 45:261–271.

Yates, F. (1981), *Sampling Methods for Censuses and Surveys* (4th ed.), Charles Griffin & Company, London.

Zhu, A. X. (1997), Measuring uncertainty in class assignment for natural resource maps under fuzzy logic. *Photogramm. Eng. Remote Sens.* 63:1195–1202.

Zhu, Z., Yang, L., Stehman, S. V., and Czaplewski, R. L. (1999), Designing an accuracy assessment for a USGS regional land cover mapping program. In *Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources* (K. Lowell and A. Jalon, Eds.), Sleeping Bear Press, Chelsea, MI, pp. 393–398.