ELSEVIER

Design and Analysis for Thematic Map Accuracy Assessment: Fundamental Principles

Stephen V. Stehman and Raymond L. Czaplewski†*

Before being used in scientific investigations and policy decisions, thematic maps constructed from remotely sensed data should be subjected to a statistically rigorous accuracy assessment. The three basic components of an accuracy assessment are: 1) the sampling design used to select the reference sample; 2) the response design used to obtain the reference land-cover classification for each sampling unit; and 3) the estimation and analysis procedures. We discuss options available for each of these components. A statistically rigorous assessment requires both a probability sampling design and statistically consistent estimators of accuracy parameters, along with a response design determined in accordance with features of the mapping and classification process such as the land-cover classification scheme, minimum mapping unit, and spatial scale of the mapping. ©*Elsevier Science Inc., 1998*

INTRODUCTION

Land-cover maps are used in numerous natural resource applications to describe the spatial distribution and pattern of land-cover, to estimate areal extent of various cover classes, or as input into habitat suitability models, land-cover change analyses, hydrological models, and risk analyses. Accuracy assessment quantifies data quality so that map users may evaluate the utility of a thematic map for their intended applications. Despite the widespread acceptance of accuracy assessment and the numerous articles published on this topic, the basic structures of a statistically rigorous accuracy assessment have not been fully described. The objective of this article is to elucidate these fundamental structures. We describe the three basic components of an accuracy assessment, the sampling design, the response design, and the estimation and analysis protocol, and we provide recommendations and general guidelines for a statistically rigorous assessment.

An accuracy assessment begins with the definition of the target population, which is the area or region represented by the land-cover map. The individual units or elements of this population are defined as pixels or polygons, depending on the map representation. A sample of units is selected from this population for accuracy assessment. Choosing the sampling unit and the sampling design are two major decisions required when planning the sampling protocol. The reference or "true" classification is obtained for each sampling unit based on interpreting aerial photography or videography, a ground visit, or a combination of these sources. The methods used to determine this reference classification are called the "response design." The response design includes procedures to collect information pertaining to the reference land-cover determination, and rules for assigning one or more reference classifications to each sampling unit. The land-cover classifications from the map are compared to the reference classifications, and the extent to which these two classifications agree is defined as map accuracy.

SELECTING THE REFERENCE SAMPLE

In this section, we discuss details of the sampling design component of the assessment. The sampling design is the protocol by which the reference sample units are selected. A probability sampling design is a key element of a statistically rigorous assessment, and several commonly used probability sampling designs are described. Implementing the sampling design also requires defining a sampling frame, along with the sampling unit which forms the basis of the accuracy assessment. Various types of frames and sampling units are described in the next two subsections.

* SUNY College of Environmental Science and Forestry, Syracuse, New York

† USDA Forest Service, Rocky Mountain Research Station, Fort Collins, Colorado

Address correspondence to Stephen Stehman, SUNY ESF, 320 Bray Hall, 1 Forestry Drive, Syracuse, NY 13210. E-mail: svstehma@mailbox.syr.edu

Received 5 August 1997; revised 26 January 1998.

REMOTE SENS. ENVIRON. 64:331–344 (1998)

655 Avenue of the Americas, New York, NY 10010

0034-4257/98/$19.00
PII S0034-4257(98)00010-8

Frames

A sampling frame consists of "the materials or devices which delimit, identify, and allow access to the elements of the target population" (Särndal et al., 1992, p. 9): the two types are "list frames" and "area frames." A list frame consists of a list of all sampling units, for example, either pixels or mapped polygons, in the target region. The sample is selected directly from this list of sampling units. An area frame provides a map or description of the population boundaries. The sampling protocol used with an area frame is based on first selecting a sample of spatial locations, followed by associating a sampling unit with each sampled location. Thus the actual sampling units, for example, polygons, are selected indirectly via the intermediate step of the sample of point locations. An explicit rule for associating a unique sampling unit with any spatial location within the area frame must be established. For example, a rule for associating a unique polygon with a randomly selected point location is to sample that polygon within which the random point fell. This particular area frame sampling protocol illustrates that it is not necessary to delineate all polygons in the population to obtain the sample. An area frame is preferable to a list frame when a systematic design is planned. For example, if the area frame is a map of all pixels, converting the map to a one-dimensional list frame of pixels would not only be unnecessary work, it would lose much of the spatial structure important for systematic sampling. Area frames better retain the spatial features of the population.

Sampling Units

The sampling unit (e.g., 0.1 ha pixel, 10 ha polygon, 1000 ha circular plot) is the fundamental unit on which the accuracy assessment is based; it is the link between a spatial location on the map and the corresponding spatial location on the earth. The response design is applied to each sampling unit to obtain the reference land-cover classification, and the comparison of the map and reference classifications is conducted on the scale of a sampling unit. For example, if a pixel is chosen as the sampling unit, the reference land-cover classification is obtained for each pixel (as represented on the earth) and compared to the corresponding map pixel. If the sampling unit is a point, the correspondence is between the classification provided by the map at that point, and the reference classification associated with the same point location on the earth.

The two types of sampling units are points and areal units. Points have no areal extent, whereas areal units possess two-dimensional spatial coverage. Pixels and polygons are examples of areal sampling units that are directly associated with mapped land-cover features. But an areal sampling unit can also be defined without reference to land-cover features of the map or ground. For example, a 1 ha areal sampling unit may encompass many pixels, each having a different land-cover classification, or span portions of several different land-cover polygons.

The choice of sampling unit is not necessarily fixed by the map representation. For example, a polygon sampling unit may be employed for a pixel-based map representation, or a point sampling unit rather than a polygon unit may be selected for use with a polygon-based map representation. The sampling unit must be defined prior to specifying the sampling and response designs, and several sampling and response design options will be available for any choice of sampling unit.

The distinction between the sampling unit and the attribute or observation recorded on that sampling unit is important. The sampling unit is just a location (point) or area in space, whereas the observations taken on the sampling unit are determined by the response design. The sampling unit can be defined without specifying what will be observed on that unit; thus no assumption about homogeneity of land-cover for the sampling unit is necessary. It is possible to use features of the map or ground to define the sampling unit. For example, land-cover polygons displayed on the classified image may be defined as sampling units. But these polygons simply determine the spatial boundary of the sampling unit, and the validity of the boundary does not require that the actual land cover within the polygon be homogeneous. That any specified spatial region (e.g., 1 ha, 10 ha, or 1 km^2 plot) can be defined as the sampling unit illustrates the independence of the sampling unit definition from the characteristics of the land cover that might be found within that sampling unit.

Point Sampling Units

The distinction between point and areal sampling units is that the statistical population associated with a point sampling unit is viewed as continuous, rather than partitioned into discrete spatial units such as pixels or polygons. A continuous population perspective avoids the difficulty of interpreting the representation or support of an individual pixel (Moisen et al., 1994). When the sampling unit is a point, the reference land-cover classification is still determined via the response design protocol. The response design may evaluate a spatial extent larger than just the point location to obtain the reference classification at that point, but the comparison of the map and reference classifications remains on a per point (sampling unit) basis. Point sampling units are usually selected from an area frame. Probability sampling concepts still apply to point sampling, and designs such as unrestricted random, stratified random, and systematic sampling are available. However, the continuous population perspective leads to some different sampling issues from those encountered in the finite population sampling framework in which accuracy assessment problems are usually treated. Some of

the statistical details are reviewed by Stehman and Overton (1996), but we will not pursue those issues further in this article.

Areal Sampling Units

The three primary areal sampling units are pixels, polygons, and fixed-area plots. Each of these sampling units partitions the population into a finite number of discrete units. Both pixels and polygons correspond to structures used in geographic information systems to represent land cover, whereas our definition of fixed-area plots does not require this correspondence. Pixels are defined by the land-cover representation of the map itself, and are usually uniform in shape and size. Pixels representing small areas (e.g., 30 m pixel) are related to point sampling units, but because pixels still possess some areal extent, they partition the mapped population into a finite, though large, number of sampling units. Larger pixels, such as the 1 km^2 pixels of AVHRR, are more closely related to the fixed-area plot sampling units defined subsequently than to point sampling units.

A polygon sampling unit is initially conceptualized as an area of homogeneous land cover displayed on the classified image (digital polygon), or identified on Earth from aerial photography or videography (photointerpreted polygon). Polygon sampling units are usually irregular in shape and differ in size. Digital polygons may be organized into a list frame or maintained in an area frame representation, whereas, for practical reasons, photointerpreted polygons are represented by an area frame. As described in the subsection on "Frames," the area frame representation of photointerpreted polygons requires delineating only those polygons identified as part of the sample, and not all polygons in the population.

Fixed-area plot sampling units are usually regular in shape, and cover some predetermined areal extent. Examples of fixed-area plot sampling units include a maplet (Stoms, 1996), defined as a high resolution map of a small geographic area (Chrisman, 1991), a video frame, an aerial photograph, and a 1 ha plot. We distinguish fixed-area plots from pixels and polygons by not restricting the fixed-area units to correspond to a land-cover structure of the map such as a pixel or digital polygon, or to a land-cover structure identified on the earth such as a photointerpreted polygon. Although in reality pixels and polygons are special cases of fixed-area plot sampling units, we distinguish these three types of areal units to focus more easily on features of each.

Selecting the Sampling Unit

No consensus exists on which sampling unit is best, and it is unlikely that any one sampling unit is optimal for all applications. Differences in project objectives, characteristics of the landscape, features of the mapping process, and practical constraints guide the choice of the sampling unit. The diversity of sampling units employed in accuracy assessment is illustrated in Table 1. Because the sampling unit is not required to match the map representation of land cover, arguments about how to best represent land cover are peripheral to the choice of a sampling unit for accuracy assessment, although these issues are still critical to the mapping effort itself (cf. Fisher, 1997). Accuracy assessment begins after a decision on the representation of land cover has been reached.

The choice of sampling unit usually represents a compromise among various benefits and costs associated with each type of sampling unit. Janssen and van der Wel (1994, p. 422) and Franklin et al. (1991) advocate using pixels as the basis of an accuracy assessment, the former arguing that "remote sensing data should be considered to be 'point-sampled' data, in which the points possess a certain spatial extent." Janssen and van der Wel (1994) further state that individual pixels are the most appropriate sampling unit for a pixel-based classification, but suggest "cluster-based sampling" when spatial smoothers have been applied, and when accessibility to terrain is poor.

Homogeneous land-cover polygons (as identified on the map) have an appealing convenient structure for the sampling unit and a direct correspondence to the land-cover representation displayed by the map. A disadvantage of using map polygons as sampling units is that the sampling units are now inseparably bound to a particular map. If subsequently this map is updated, for example, after a revised classification is developed to improve accuracy, the original polygon sampling units are still valid for the assessment, but they may no longer correspond to land-cover polygons of the revised map. A similar issue arises in a change detection accuracy assessment. A mapped land-cover polygon used to define a sampling unit at one point in time may not exist at a later point in time. Hierarchical land-cover classification schemes present a related problem when defining the sampling unit based on a map polygon. At which level in the classification hierarchy should the polygons be identified? How is this sampling unit then used when assessing accuracy at a different level of the classification scheme in which the land-cover polygons differ from those identified at another level in the hierarchical scheme?

Fixed-area plots defined independently of land-cover polygons retain their identity under map revisions and over time. The disadvantage is that these units do not correspond directly to landcover polygons, either of the map or the ground. For example, a 1 ha areal sampling unit may include portions of several different land-cover polygons, or contain several smaller polygons within the sampling unit. The nonsite specific character of large fixed-area plots may result in an assessment that is too coarse for some uses of the data, such as when the small-scale spatial distribution of land cover is more important to the objectives than regional estimates of land-cover area proportions.

Table 1. Sampling Units Employed or Recommended for Various Accuracy Assessment Projects[a]

Project	*Sampling Unit*
Bauer et al. (1994)	88-Acre unit (psu), pixel (ssu)
Cibula and Nyquist (1987)	3×3 pixel block
Clerke et al. (1996)	400 ha (psu), polygon (ssu)
Conese and Maselli (1992)	Pixel
Congalton et al. (1993)	Polygon (aerial photograph)
Dicks and Lo (1990)	5-Acre grid cell
Edwards et al. (1998)	1 ha plot within psu
Felix and Binney (1989)	Polygon (map)
Fenstermaker (1991)	3×3 pixel block
Fiorella and Ripple (1993)	Pixel
Fitzpatrick-Lins (1981)	Point
Franklin et al. (1991)	3×3 pixel block
Fung and LeDrew (1988)	Pixel
George (1986)	Polygon (map)
Hord and Brooner (1976)	1-Acre plot
Knick et al. (1997)	Pixel
Lauver and Whistler (1993)	Polygon (grassland)
Martin (1989)	3×3 cluster for psu, individual pixel for ssu
Martin and Howarth (1989)	3×3 pixel block (psu), pixel (ssu)
McGwire et al. (1996)	Pixel
Riley et al. (1997)	Pixel
San Miguel-Ayanz and Biging (1996)	4×4 pixel block
San Miguel-Ayanz and Biging (1997)	4×4 pixel block for TM, 6×6 pixel block for SPOT
Senseman et al. (1995)	Pixel
Stenback and Congalton (1990)	3×3 pixel block
Stoms (1996)	Maplet
Todd et al. (1980)	9×9 cluster for psu, 3×3 pixel block for ssu
Vujakovic (1987)	2×2 pixel block
Walsh et al. (1987)	2.5- and 10-acre cells
Warren et al. (1990)	Polygon
Wickware and Howarth (1981)	Pixel
Zhu et al. (1996)	1 km^2
Zhuang et al. (1995)	Pixel

[a] psu=primary sampling unit; ssu=secondary sampling unit.

Confounding of classification and location error is a troublesome problem in accuracy assessment, and it is not clear which sampling unit to choose on the basis of sensitivity to location error. To avoid location error, the reference sample is sometimes restricted to polygon interiors or to pixels within homogeneous blocks. In such cases, the accuracy assessment represents a portion of the map, which can be a small proportion of the total area if most polygons or pixel blocks of homogeneous land cover are small. Restricting the assessment to homogeneous areas is not a recommended strategy because of the optimistic accuracy results that typically arise (Hammond and Verbyla, 1996). Once boundaries and edges are included in the sample, location error seems equally problematic whether the sampling unit is a pixel, a polygon, or a larger area. Stehman and Czaplewski (1997) present some methods for accommodating potential effects of location error in the analysis.

In summary, the sampling unit is a structure that defines a specified point or area of space. The sampling unit is a structure we impose in the specification of the sampling design, and it is not determined by the map representation of land cover. Even if there are convenient or naturally occurring sampling units such as pixels or polygons, we are neither obligated nor prevented from selecting these units for the assessment. Choosing a sampling unit may require considering issues such as location error, minimum mapping unit, and how polygon boundaries will be treated in the assessment. Because the sampling unit is the ultimate basis for the comparison of the map and reference classifications, whatever sampling unit is chosen, it is essential that this choice be explicitly and clearly stated and acceptable to users of the thematic map.

Sampling Design

The sampling design is the protocol by which sampling units are selected into the sample. Implementing a probability sampling design contributes to a scientifically defensible accuracy assessment, and Smith (1990) argues for such designs because of their objectivity. Probability sampling is defined in terms of inclusion probabilities, which represent the probability of including a particular sampling unit in the sample. Inclusion probabilities are derived from the set of all possible samples that could result from a sampling design protocol, so they represent what we expect *prior* to choosing the actual reference

sample. Särndal et al. (1992, Section 2.4) review inclusion probabilities in more detail. Probability sampling requires that *all* inclusion probabilities be greater than zero, and the inclusion probabilities must be known for those units selected in the sample. If some sampling units have an inclusion probability of zero, the assessment does not represent the entire target region of the map. Excluding inaccessible areas or heterogeneous edges between polygons is an example of assigning sampling units an inclusion probability of zero. Requiring the inclusion probabilities to be known is necessary so that statistically valid (i.e., consistent) estimates can be computed.

Simple random, stratified random, cluster, and systematic sampling are all probability sampling designs. When using such designs in practice, the inclusion probabilities do not have to be computed explicitly because they are already taken into account in the standard estimation formulas. But if a new or nonstandard sampling protocol is constructed, then the investigators must specify the inclusion probabilities. The inclusion probabilities determine the weight attached to each sampling unit in the estimation formulas, and if the inclusion probabilities are unknown, so are the estimation weights. A good rule to apply when planning an accuracy assessment is that if the sampling protocol cannot be identified as a standard probability sampling design and the project planners are unable to specify the nonzero inclusion probabilities, the proposed design should be discarded.

COMPARISONS OF COMMON PROBABILITY SAMPLING DESIGNS

Basic probability sampling designs are constructed from simple random and systematic selection protocols, and structures imposed on the population such as strata and clusters. Simple random and systematic selection protocols may be applied to a population with or without strata or clusters. Within the class of stratified designs, stratified random sampling in which a simple random sample is obtained in each stratum is most commonly employed. But a systematic selection protocol may be employed to sample within strata, and it is even possible to have some strata sampled systematically and others sampled via simple random sampling, all within the same stratified design. The class of cluster sampling designs includes simple random or systematic selection of clusters, and also two-stage cluster sampling in which the units within each sampled cluster are themselves sampled. Systematic sampling using a regular grid and stratified systematic unaligned sampling are options within the class of systematic designs.

Simple random, systematic, stratified systematic unaligned sampling, and one-stage cluster sampling, with the clusters selected via simple random or systematic sampling, are all equal probability sampling designs. Stratified sampling with proportional allocation also results in equal inclusion probabilities, but stratified sampling with either equal or optimal allocation usually leads to different inclusion probabilities for the sampling units in different strata. If polygons are sampled by selecting those polygons in which randomly chosen point locations fall, larger polygons have a higher probability of being "hit" by a random point, and therefore have higher inclusion probabilities. Unequal inclusion probabilities create no difficulties as long as they are known and accounted for in the estimation formulas, but equal probability designs possess the advantage of simpler analysis.

Of the two basic selection protocols, simple random and systematic, systematic is often easier to implement, particularly when an area frame is employed. Because systematic sampling produces a spatially well-distributed sample, it usually results in better precision relative to simple random sampling. The choice between simple random or systematic sampling is also affected by the importance of unbiased variance estimation to assessment objectives. Systematic designs, including stratified systematic unaligned sampling, do not permit unbiased estimation of variance, and the true variance is usually overestimated from the sample data. It is critical to recognize that the concern with systematic sampling is not unbiased estimation of the accuracy parameters themselves, but rather unbiased estimation of the *uncertainty* or *variability* of these estimates (Stehman, 1992).

Stratification is a frequently employed design structure with geography and mapped land-cover class being two of the common stratification attributes. Geographic stratification can be used to distribute sampling effort evenly among administrative regions or ecoregions, or to sample accessible areas with higher probability than expensive, but low-priority, inaccessible regions. Stratifying by mapped land-cover classes may ensure that a specified sample size is obtained in each mapped class, including those rare classes that would not be prevalent in a simple random or systematic sample without stratification. A disadvantage of stratifying by mapped land-cover class is that it locks the assessment into the map version used to form the strata. If this map is subsequently revised or the land-cover classification scheme changed, the original strata are still valid, but they no longer correspond to the land-cover classes of the revised map. Stratifying by the mapped land-cover classes requires the map to be available prior to selecting the sample, and this may cause a delay between when the imagery is obtained and when the reference data are collected.

Cluster sampling employs two sizes of sampling unit. The clusters themselves are the primary sampling units (psu), and the units making up the cluster are the secondary sampling units (ssu). A variety of structures have been used to form clusters. Commonly the cluster is a block of pixels, for example, a 3×3 or 5×5 block, but clusters may also be formed by grouping pixels in a linear arrangement (Edwards et al., 1998). Another form of cluster sampling is to use "cluster plots." In this ap-

proach, the cluster consists of a centrally located ssu surrounded by other secondary units arranged in some specified pattern. For example, the center unit may be a 100 m^2 plot, and four other 100 m^2 plots in the cluster are located a specified distance from the central unit along the four compass directions. The ssu's are not contiguous in this version of cluster sampling.

The ssu is the ultimate basis of the comparison between the map and reference classifications. That is, the reference classification should be obtained for *each* ssu (usually a pixel) within the psu (the cluster of pixels), and the comparison of the map and reference classifications is then made for each ssu. This is not always the protocol followed. For example, sometimes a block of homogeneous pixels is used as the selection unit, but only the center pixel is used for the assessment. Because the comparison of the map and reference classifications is based on only the center pixel, the sampling unit is in reality just this center pixel, not the block of pixels, so the design should not be considered cluster sampling. In other cases, the comparison of the reference and map classifications is made at the spatial scale of the block of pixels, not on per pixel agreement. That is, the majority land-cover class from a mapped 3×3 block of pixels may be compared to a single reference classification combining information over all nine reference pixels in the block. This is not truly a cluster sampling design because the sampling unit is not a pixel within a cluster, but rather the 3×3 block of pixels. The map and reference comparison is not on a per pixel basis, as is required to define a pixel sampling unit, but on a per block basis, making the block the sampling unit. Regarding the 3×3 pixel cluster as the sampling unit would be appropriate if a 3×3 spatial smoother had been applied to the entire map.

Cluster sampling is motivated by the potential reduction in the sampling cost per ssu (Moisen et al., 1994). For example, it is less expensive to sample all nine pixels within a psu defined as 3×3 block than it is to sample nine pixels located at random throughout the study area. The cost reduction achieved by cluster sampling must be sufficiently large to compensate for the loss of information per sampling unit (ssu) attributable to the intracluster spatial correlation among ssu's. Moisen et al. (1994) provide guidelines illustrating combinations of cluster size and intracluster correlation favorable to employing clusters in accuracy assessment. A disadvantage of cluster sampling is that the standard error formulae are more complex than those for simple random sampling because it is necessary to account for the lack of independence among the secondary sampling units within a cluster (Czaplewski, 1994; Stehman, 1997a).

Two-stage cluster sampling is often used to provide spatial control over the sample to reduce costs. In this design, large psu's, for example aerial photographs or 1:24,000 quad maps, are selected at the first stage of sampling, and then a subsample of the ssu's (e.g., 1 km^2 plots) within each psu is obtained. If field visits are necessary, then most of the travel effort is concentrated within the spatially limited area defined by each psu. Two-stage cluster sampling may also be employed to diminish the variance inflation effect a high positive intracluster correlation has on one-stage cluster sampling. Edwards et al. (1998) and Zhu et al. (1996) provide good examples of this design.

Nonprobability Sampling

Unfortunately, examples of nonprobability sampling are common in accuracy assessment applications. Selecting reference locations by purposeful, convenient, or haphazard procedures does not provide the structure to determine the inclusion probabilities for each sampling unit. Such designs, therefore, are not probability samples. Purposefully selecting training data for a supervised classification is a good example of a nonprobability sample. Such samples are acceptable for developing the land-cover classification, but often have limited use for accuracy assessment because the necessary probability foundation to permit generalization from the sample data to accuracy of the full population is lacking.

Selecting the reference sample from conveniently accessible sites or available aerial photography suffers from the same problem. It is virtually impossible to assert with any confidence that these convenient sources of data have the same attributes as the entire region. We may *assume* this to be the case, but this assumption cannot be scientifically defended. Readily accessible locations or available aerial photography may represent a valid subarea of the mapped region, but it is not statistically justified to infer accuracy of the entire region from this subset. Stratified probability sampling based on maps of accessibility zones can diminish the pragmatic problems of inaccessibility without having to resort to nonprobability sampling and the associated problems with defending untestable assumptions.

Nonprobability sampling also results from purposeful selection of flight lines for collecting reference data using videography. Even when a subsample of video frames is used for the actual reference data, if the original flight lines were not selected according to a probability sampling protocol, the design cannot be classified as a probability sample of the full region, although it may serve as a probability sample of the subregion covered by available videography. It is possible to obtain useful information from nonprobability samples, but the limitations of such data should be recognized.

RESPONSE DESIGN

The response design is the protocol for determining the reference land-cover classification of a sampling unit. Conceptually it is useful to separate the response design

into two components, the evaluation protocol, which consists of the procedures used to collect information contributing to the reference classification determination, and the labeling protocol, which assigns a land-cover classification to the sampling unit based on the information obtained from the evaluation protocol. The resulting reference classification must have high accuracy for a valid assessment (Congalton, 1991). Congalton and Green (1993), Hammond and Verbyla (1996), and Verbyla and Hammond (1995) describe some of the difficulties inherent in obtaining accurate reference classifications.

We emphasize again that the sampling unit serves as the basic unit of comparison between the map classification and the reference classification. Although pixel and polygon sampling units are often assumed to consist of a single land-cover class, this homogeneity of land-cover within the sampling unit is an appealing, but not necessary feature. The response design can accommodate sampling units possessing homogeneous or heterogeneous land cover. Because of the possibility that any areal sampling unit, even a small pixel, may consist of more than one land-cover type, assessments based on areal units are always to some extent non-site-specific. The larger the sampling unit, the more the assessment takes on this non-site-specific character. Merchant et al. (1993) present an excellent discussion of issues related to non-site-specific assessments, particularly as they apply to AVHRR pixels.

Evaluation Protocol

The first step in developing the response design is to choose the spatial support region on which the reference land-cover evaluation will be based. Atkinson and Curran (1995, p. 768) define spatial support as "the size, geometry and orientation of the space on which an observation is defined." For example, if the sampling unit is a point, the evaluation need not be limited only to what the evaluator observes at that point location. Rather, the evaluation may be based on a more general landscape view encompassing a larger surrounding area, say 100 m^2, 1 ha, or 1 km^2. The response design also includes specifying the area and shape of the support region, both possibly depending on the type of land cover. Linear features such as utility corridors or stream riparian zones may be evaluated differently from forest stands or agricultural fields. The evaluation protocol may allow support regions from different sampling units to overlap.

A spatial support region defined for an areal sampling unit may or may not be the areal unit itself. For example, a 30 m pixel may be assigned a support region of 1 ha, whereas a 1 km^2 AVHRR pixel may be assigned a support region matching the size of the pixel. Fisher (1997) discusses some of the difficulties in defining the support area of a pixel, and these same issues apply to both large and small pixels and fixed-area plot sampling units. The spatial support of a polygon sampling unit will usually just be the polygon itself. For example, if the sampling unit is defined by the boundary of a mapped land-cover polygon, this same boundary may define the support region in the evaluation protocol. Whatever support region is specified, the eventual reference classification applies to the sampling unit, not the spatial support region.

Once the support region has been identified, numerous options are available to determine the reference classification. In some cases, the evaluator may visually scan the support region and record qualitative observations contributing to an eventual classification of the sampling unit. In other cases, the evaluation protocol may specify recording species composition, canopy closure, or distribution of tree sizes, or require other quantitative data needed to distinguish among land-cover classes or to characterize the land cover of the sampling unit. The evaluation protocol should conform to the users' concept of error-free classification; any compromises should be agreeable to users.

The evaluation protocol may include sampling within the areal unit. This subsampling within the response design contributes to the land-cover classification recorded on a sampling unit, but it is not part of the structure required for the sampling design and analysis components. Line transects, quadrats, or gridded point samples are candidate response design sampling methods for estimating quantitative characteristics that contribute to the land-cover classification of a sampling unit. The response design sampling also provides information on within-pixel or within-polygon heterogeneity. This information may be relevant to the subsequent labeling protocol, or to characterize heterogeneity within a particular land-cover class. However, the primary objective of the response design is to obtain information pertinent to identifying a reference land-cover label for each sampling unit. Ground data are sometimes collected for objectives other than land-cover determination. Curran and Williamson (1986), McGwire et al. (1993), and Steven (1987) discuss issues related to quantitative characterization of ground plots for features such as reflectance ratios, green leaf area index, and biomass.

Labeling Protocol

The labeling protocol assigns the reference classification (or classifications) to the sampling unit based on the information obtained from the evaluation protocol. At the most basic level, the reference sampling unit is labeled as one and only one land-cover class. This primary class labeling suffers from the potential problem that a sampling unit may consist of several different land-cover classes, or represent a transition or mixed class not easily identified as a single cover type. Because it is not always possible or desirable to label the sampling unit as a single land-cover class, the labeling protocol may specify re-

cording both a primary and secondary land-cover class (cf. Edwards et al., 1998).

Similar concerns lead to the "fuzzy" classification approach in which the evaluation protocol provides a qualitative assessment of class membership for each possible land-cover category (Gopal and Woodcock, 1994). For example, a linguistic scale employed in a fuzzy classification may range from "absolutely wrong" to "absolutely right." Once the fuzzy evaluation has been obtained, a labeling protocol must still be applied. Gopal and Woodcock (1994) define RIGHT and MAX operators as two options for assigning a label. The RIGHT operator assigns the label of any land-cover class that scored above a certain level in the fuzzy evaluation, so that it is possible that several land-cover classes would be assigned to a sampling unit. The MAX operator assigns the label of the land-cover class having the highest evaluation score. The information from a fuzzy evaluation protocol could also be used to assign primary and secondary land-cover classes to the sample units.

If the evaluation protocol generates quantitative land-cover data, such as area proportions for each land-cover class present in an areal sampling unit or spatial support region, a quantitative labeling protocol becomes an option. That is, the reference classification for a sampling unit may be a vector of area proportions, for example, 0.2 Old Growth Forest, 0.3 Forest (not Old Growth), and 0.5 Non-Forest. This protocol is one way the response design can be implemented to accommodate heterogeneity of land cover within a sampling unit. The information obtained from a quantitative labeling protocol can also be summarized to provide primary and/or secondary land-cover classes. For the example provided, the primary label would be Non-Forest, and the secondary label would be Forest.

Selecting a Response Design

The response design is chosen depending on the procedure for assessing agreement (e.g., primary, fuzzy, or quantitative), the sampling unit, and the information needed to ascertain the reference land-cover classification. Features of the mapping process such as the classification scheme, minimum mapping unit, and spatial scale also influence the response design choice. Probability sampling may be employed within the response design itself to ensure objectivity, and to distinguish among land-cover classes defined by quantitative characteristics. But if the true land cover of the sampling unit can be obtained better by nonprobability sampling methods, then that option may be exercised in the response design component. Because the response design addresses the fundamental question of how to characterize the land cover of a parcel of ground, it is selected in adherence with prevailing conventions of land-cover classification and the requirements of those using the thematic map. In accuracy assessment, as is typical of most sampling investigations, the attributes measured on a sampling unit are decided by the subject matter specialists. Thus the response design requires input from scientists and analysts having a clear understanding of the land-cover classification scheme being used in the mapping project. Because of the interpretive nature inherent in determining land-cover classification, the response design may also require a reliability or quality control component to evaluate the repeatability and even the accuracy of the reference land-cover classifications themselves.

The land-cover map is not always the final product from the users' perspectives. These maps are often input to predictive models within geographic information systems. Accuracy assessment of model predictions is another important objective: How well do model predictions based on mapped data agree with predictions based on reference data? It is entirely possible that a map with poor thematic accuracy can produce acceptably accurate model predictions if the model is not sensitive to the types of categorical confusion within the map. When feasible, the response design should also accommodate collecting data necessary to evaluate important prediction models and other analyses in which users will incorporate the land-cover mapping information. Map users must determine if the protocols for the reference classifications conform to their needs. For example, if a user's intended analyses presume quantitative field protocols, but qualitative field observations are used for the reference protocol, then the accuracy assessment will not present a relevant evaluation of data quality for that user's needs.

ANALYSIS AND ESTIMATION

The analysis and estimation protocols applied to the reference sample data constitute the third main component of an accuracy assessment. An error matrix (Table 2) effectively summarizes the key information obtained from

Table 2. Population Error Matrix for a Land-Cover Scheme of q Classes[a]

		Reference				
		1	*2*	…	*q*	
Map	1	p_{11}	p_{12}	…	p_{1q}	p_{1+}
	2	p_{21}	p_{22}	…	p_{2q}	p_{2+}
	⋮	⋮	⋮	…	⋮	⋮
	q	p_{q1}	p_{q2}	…	p_{qq}	p_{q+}
		p_{+1}	p_{+2}	…	p_{+q}	

[a] Notes: 1) p_{ij} is the proportion of area in mapped land-cover class i and reference land-cover class j; 2) $p_{i+}=\sum_{j=1}^{q} p_{ij}$ is the proportion of area mapped in land-cover class i; and 3) $p_{+j}=\sum_{i=1}^{q} p_{ij}$ is the true proportion of area in land-cover class j.

the sampling and response designs. The error matrix represents a contingency table in which the diagonal entries represent correct classifications, or agreement between the map and reference data, and the off-diagonal entries represent misclassifications, or lack of agreement between the map and reference data. Typically, the error matrix summarizes results comparing the primary reference class label to the map land-cover class for the sampling unit, and the results presented in this section focus on this situation. However, error matrices can also be constructed in which other labeling schemes are used in the response design. For example, agreement could be defined as a match between the map classification and either the primary or secondary class. Gopal and Woodcock (1994) describe methods to summarize results for a fuzzy classification, and these methods are illustrated, for example, in Knick et al. (1997). Zhu et al. (1996) demonstrate how a quantitative reference labeling scheme can be displayed and summarized.

In Table 2, the column labels represent the reference classifications, and the row labels represent the map classifications. The cell proportions, p_{ij} form the basis of the error matrix summary. These proportions may be derived from pixel or polygon counts, or measurement of areas, depending on the user's preference. If the assessment is based on equal area pixels, then p_{ij} is the same for counts and areas. A polygon-based assessment results in a difference between p_{ij} for polygon counts and p_{ij} for polygon areas. For simplicity, our discussion will focus on p_{ij} as representing proportion of area, with p_{ij} interpreted as the proportion of area classified as land-cover category i by the map and category j by the reference data. The row sum, $p_{i+}=\sum_{k=1}^{q} p_{ik}$, is the proportion of area mapped as land-cover class i, and the column sum, $p_{+j}=\sum_{k=1}^{q} p_{kj}$, is the true proportion of area in land-cover class j (q=number of land-cover classes). In practice, the p_{ij} must be estimated from the sample data. These estimates $\hat{p}_{ij}$ are then used to construct estimates of accuracy parameters.

Accuracy Parameters

Various summary measures are derived from the error matrix to describe accuracy. We focus on population parameters which represent well-defined probabilities of either correct classifications or various misclassifications. Numerous other accuracy parameters not directly interpretable in this probability framework have been proposed, but it is sometimes difficult to interpret how these parameters are related to features of the actual map being assessed (Stehman, 1997b).

A core set of accuracy parameters that can be interpreted as probabilities defined for the map being assessed includes the following:

1. Overall proportion of area correctly classified,

$$P_c=\sum_{k=1}^{q} p_{kk}, \tag{1}$$

   which represents the probability that a randomly selected point location is classified correctly by the map.
2. User's accuracy for land-cover class i, the conditional probability that a randomly selected point classified as category i by the map is classified as category i by the reference data,

$$P_{Ui}=p_{ii}/p_{i+}. \tag{2}$$

3. Producer's accuracy for land-cover class j, the conditional probability that a randomly selected point classified as category j by the reference data is classified as category j by the map,

$$P_{Aj}=p_{jj}/p_{+j}. \tag{3}$$

4. Probability of a commission error, which is the conditional probability that a randomly selected point classified as category i by the map is classified as category k by the reference data,

$$p_{ik}/p_{i+}. \tag{4}$$

5. Probability of an omission error, which is the conditional probability that a randomly selected point classified as category j by the reference data is classified as category k by the map,

$$p_{kj}/p_{+j}. \tag{5}$$

Estimating Accuracy Parameters

Obtaining $\hat{p}_{ij}$ is the first step in the analysis protocol. The probability sampling character of the sampling design is critical here because estimating $\hat{p}_{ij}$ must incorporate the known inclusion probabilities for the design used. For example, for a simple random sample of n pixels from N pixels in a map, $\hat{p}_{ij}=n_{ij}/n$, where n_{ij} is the number of reference sample pixels classified as map category i and reference category j. For a stratified random sample based on the mapped land-cover classes as strata, $\hat{p}_{ij}=(n_{ij}/n_{i+})(N_{i+}/N)$, where n_{i+} and N_{i+} are the sample and population sizes in stratum i. To estimate other accuracy parameters, $\hat{p}_{ij}$ is substituted for p_{ij} in the formula for the accuracy parameter. For example, substituting $\hat{p}_{ij}$ for p_{ij} in the formula for producer's accuracy,

$$P_{Aj}=p_{jj}/p_{+j}=p_{jj}\Big/\sum_{k=1}^{q} p_{kj}, \tag{6}$$

leads to the estimated producer's accuracy,

$$\hat{P}_{Aj}=\hat{p}_{jj}/\hat{p}_{+j}=\hat{p}_{jj}\Big/\sum_{k=1}^{q} \hat{p}_{kj}. \tag{7}$$

Because it is easier to combine proportions properly to estimate parameters of interest, we recommend re-

porting the error matrix in terms of proportions ($\hat{p}_{ij}$) rather than counts (n_{ij}).

The estimation approach described results in consistent estimators of the parameters of interest (Särndal et al., 1992, Section 5.3; Stehman, 1995). An estimator is defined to be consistent if the estimator is the same as the population parameter when the sample size is increased to where it matches the population size (Cochran, 1977). In practical terms, consistency ensures that we are estimating the targeted parameter of the population of interest. Inconsistent estimators arise when the estimation formulas do not match the sampling design, such as when simple random sampling formulas are used with an equally allocated stratified design. Normalizing an error matrix also leads to inconsistent estimates. The motivation for normalizing an error matrix is to create a standardization that allows for comparing error matrices. Rather than normalizing, employing conditional probabilities based on either the row or column marginal proportions (p_{i+} or p_{+j}) provides a more interpretable and defensible standardization. For example, probabilities conditioned on the row marginal proportions represent user's accuracy and commission error probabilities; within each row, these conditional probabilities sum to 1, which may be viewed as a standardization eliminating differences among row marginal proportions. Similar characteristics exist for producer's accuracy and omission error probabilities, which represent probabilities conditioned on the column marginal proportions.

Comparisons between two error matrices are readily accomplished using the conditional probabilities represented by the parameters (2)–(5) presented in the previous subsection. Normalizing an error matrix may be viewed as an attempt to condition simultaneously on *both* row and column marginal proportions. It is difficult to interpret what the probabilities resulting from such a simultaneous conditioning represent in terms of the real population being assessed. Consequently, describing characteristics of the hypothetical population represented by a normalized error matrix contributes little interpretive value to the accuracy assessment.

Variance estimation is an important feature of the analysis component. For estimating each of the parameters in Eqs. (2)–(5), a different variance estimator formula arises for each different sampling design. We do not catalog these many variance estimator formulae here. Stehman (1995) outlines an approach to variance estimation in accuracy assessment, and the general estimation framework of Czaplewski (1994; 1998) provides a more detailed treatment of variance estimator formulae.

Statistical Software

Most commonly available statistical software and spatial analysis programs will not compute estimates for the parameters of interest in accuracy assessment, or if estimates are provided, standard errors will be available only if the sampling design is simple random sampling. Most of the estimation formulas needed for accuracy assessment can be programmed into standard spreadsheet packages. Specialized software exists for estimating parameters and standard errors for sampling designs more complex than simple random sampling [see Lepkowski and Bowles (1996) for a review], but this software may not be readily available or familiar to users. Williams and Beach (1995) have developed a general estimation program specifically for accuracy assessment based on the results reported in Czaplewski (1992; 1994; 1998).

GENERAL RECOMMENDATIONS

Accuracy assessments typically have multiple users and objectives leading to interest in a variety of accuracy parameters and subregions of the mapped area. In those rare cases in which the accuracy assessment objectives are limited, specialized designs can be tailored to meet these few objectives. For example, if the objective is to evaluate contract compliance for image classification, stratified sampling can be employed to provide adequate sample sizes in each mapped land-cover class to determine if the map satisfies the contractual accuracy requirement specified for each class. But most land-cover mapping programs have multiple users, as well as unspecified potential future applications and users. The need to satisfy multiple objectives motivates selecting a simple, general purpose sampling design. Simplicity is a key criterion because simple designs are easier to implement properly in the field and to analyze, and they are more likely to provide adequate information for a broad variety of objectives. Simple designs are also easier to understand, so the accuracy assessment data are more likely to be used correctly, even by future users who may not be familiar with the planning and details of the design.

A disadvantage of a broadly adequate, simple design is that it will be less effective for any single objective relative to a design tailored specifically for that objective. For example, if a rare class is critical to the success of a mapping project, a specialized, separate design can be added to augment the sample size in the rare class which will likely not be well represented in a simple, general sampling design (Aronoff, 1982; Congalton, 1991; Edwards et al., 1998; Fitzpatrick-Lins, 1981). The supplemental sample should follow probability sampling protocols, and if the augmented sample data are combined with the original sample, the combined-sample estimators must still satisfy the consistency criterion. The details for treating an augmented sample in a statistically rigorous manner remain to be documented.

The high cost of obtaining reference data motivates an attempt to increase efficiency, and consequently reduce costs, by employing design structures such as cluster plots and accessibility strata. What is less recognized

is that often the estimation part of the sampling strategy can be used to advantage to reduce variability while incurring little or no additional sampling cost. Poststratified and regression estimators are examples in which additional sampling is not needed. Initially proposed by Card (1982) for application in accuracy assessment, poststratification can achieve modest reductions in standard error (3–10%, in most cases) for estimating P_c and producer's accuracies using information already available in the assessment (Stehman, 1996b), the only added "cost" being that associated with using a more complex estimator. Regression estimation requires additional information, but not necessarily additional field sampling, and gains in precision are possible for less cost than would be required to obtain additional reference sample units (Stehman, 1996a). Gaining efficiency by employing more sophisticated estimators should routinely be considered as a practical, cost-saving measure in accuracy assessment. Czaplewski (1992; 1998) establishes a very general estimation framework based on a multivariate composite estimator for using auxiliary information to improve precision of accuracy estimates.

If a poor design is implemented, collecting new data is prohibitively expensive, and sometimes impossible if too much time has passed since the imagery was obtained. Reanalyzing data, even long after the reference sample has been collected, is relatively inexpensive. Flexibility afforded by the design for later reanalysis and sample augmentation is thus a relevant design criterion. Simple designs are more amenable to more complex analysis techniques such as regression and poststratified estimators and to design modifications such as supplementing the sample for rare classes.

Although practical considerations play a prominent role in accuracy assessment planning, these considerations should not lead to use of inefficient or incorrect sampling designs and analyses. Practical limitations do determine what realistically can be expected of statistical methods, and this should focus accuracy assessment planning on the priority objectives of the mapping project. If all objectives cannot be addressed well, the sampling strategy must be constructed so that critical issues are addressed adequately. Secondary objectives may, by necessity, not receive adequate sampling resources. A practical accuracy assessment sampling strategy often represents a compromise, with the overall design goal being adequacy for all critical objectives, not optimality for any single objective.

SUMMARY

If scientifically sound management and policy decisions are to incorporate information available from land-cover maps, these maps should be accompanied by a statistically rigorous, defensible accuracy assessment. The primary components of the accuracy assessment are the sampling design, response design, and estimation and analysis protocols. A great deal of flexibility is available in selecting among the options for each of these components, and decisions should be based on the strengths and weaknesses of each option to meet project objectives and practical constraints.

The sampling design should be a probability sampling design to ensure a rigorous statistical foundation for inference. If a probability sample from the entire map region is not feasible, then a probability sample from some portion of the region is a better alternative than foregoing probability sampling entirely. Deviations from probability sampling protocol are sometimes unavoidable because of practical constraints. When such deviations occur, additional data quality information should be presented to indicate how these deviations may affect the results of the accuracy assessment (Stehman and Czaplewski, 1997). Whatever sampling design is chosen, the selected sample units should be displayed geographically so that the spatial distribution of the sample is apparent for diagnostic and descriptive purposes. In the planning and description of the sampling design, it is also critical that the sampling unit and sampling design be clearly and correctly identified. This is unfortunately rarely the case (Hammond and Verbyla, 1996). Too often, ambiguous terminology such as sample "site" or sample "location" is used to describe the sampling unit. The terms "representative" and "random" sampling are often used, but these terms lack an agreed-upon, unambiguous definition and should be discarded from the lexicon of accuracy assessment. The more rigorous, well-defined classification of designs as probability and nonprobability sampling designs should be adopted.

The analysis should focus on accuracy parameters that represent probabilities of encountering certain kinds of misclassification errors or correct classifications characteristic of the mapped region of interest. Parameters such as kappa, tau, and other summary measures that cannot be interpreted in this framework should be used with caution (Stehman, 1997b). The populations of interest in accuracy assessment are real, tangible entities, so the parameters estimated should be characteristics of these real populations. This is ensured by adhering to the criterion of consistency for the estimators employed in the assessment, and by focusing on the parameters described in the Analysis and Estimation section. Because estimates from a normalized error matrix violate the consistency criterion, we discourage users from normalizing error matrices prior to estimating accuracy parameters.

If the decision is made that a land-cover mapping project will be accompanied by a statistically defensible accuracy assessment, a price must be paid to attain the statistical support desired for this assessment. While it is certainly more convenient to use available data or to restrict samples to readily accessible locations or sources, if the protocols of probability sampling and consistent es-

timation are not adhered to, the statistical, and therefore scientific foundation of the assessment is greatly eroded. We cannot carry out accuracy assessments in a nonstatistical fashion and hope to gain the scientific credibility of a statistically rigorous assessment. It is, of course, always an option to conduct an accuracy assessment that is not intended to rely on statistical support, and these assessments should be clearly identified as not being derived from a statistically based design and analysis protocol. It is obviously misleading to apply statistical techniques to an accuracy assessment without having followed the necessary protocols of sampling design and estimation.

To state that a balance must be found between statistical validity and what is practical suggests that these are conflicting features of accuracy assessment. But a statistically valid accuracy assessment need not be complex or impractical, and a practical assessment lacking statistical validity is not credible. Practicality is a characteristic of any effective sampling strategy: The key is to apply appropriate statistical tools so that both practicality and statistical rigor are satisfied. A simple sampling strategy for the assessment will often provide both the needed information and statistical rigor, and yet still be cost-effective and easy to implement. The basic structures we describe form the building blocks with which such a practical, statistically rigorous assessment can be constructed.

We thank Anthony Olsen and Scott Urquhart for suggesting the concept of a response design, and Zhiliang Zhu and the anonymous reviewers for helpful comments. This research was partially supported by cooperative agreement CR821782 between the Environmental Protection Agency and SUNY-ESF. This manuscript has not been subjected to EPA's peer and policy review, and does not necessarily reflect the views of the Agency. Additional support was provided by the Forest Inventory and Analysis Program within the USDA Forest Service.

REFERENCES

Aronoff, S. (1982), The map accuracy report: a user's view. *Photogramm. Eng. Remote Sens.* 48:1309–1312.

Atkinson, P. M., and Curran, P. J. (1995), Defining an optimal size of support for remote sensing investigations. *IEEE Trans. Geosci. Remote Sens.* 33:768–776.

Bauer, M. E., Burk, T. E., Ek, A. R., et al. (1994), Satellite inventory of Minnesota forest resources. *Photogramm. Eng. Remote Sens.* 60:287–298.

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

Chrisman, N. R. (1991), The error component in spatial data. In *Geographical Information Systems—Volume 1: Principles* (D. J. Maguire, M. F. Goodchild, and D. W. Rhind, Eds.), Longman, New York, pp. 165–174.

Cibula, W. G., and Nyquist, M. O. (1987), Use of topographic and climatological models in a geographical data base to improve Landsat MSS classification for Olympic National Park. *Photogramm. Eng. Remote Sens.* 53:67–75.

Clerke, W., Czaplewski, R. L., Campbell, J., and Fahringer, J. (1994), Assessing the accuracy of a regional land cover classification. In *Spatial Accuracy Assessment in Natural Resources and Environmental Sciences* (H. T. Mowrer, R. L. Czaplewski, and R. H. Hamre, Eds.), General Technical Report RM-GTR-277, USDA Forest Service, Fort Collins, CO, p. 508.

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

Conese, C., and Maselli, F. (1992), Use of error matrices to improve area estimates with maximum likelihood classification procedures. *Remote Sens. Environ.* 40:113-124.

Congalton, R. G. (1991), A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sens. Environ.* 37:35–46.

Congalton, R. G., and Green, K. (1993), A practical look at the sources of confusion in error matrix generation. *Photogramm. Eng. Remote Sens.* 59:641–644.

Congalton, R. G., Green, K., and Teply, J. (1993), Mapping old growth forests on national forest and park lands in the Pacific Northwest from remotely sensed data. *Photogramm. Eng. Remote Sens.* 59:529–535.

Curran, P. J., and Williamson, H. D. (1986), Sample size for ground and remotely sensed data. *Remote Sens. Environ.* 20:31–41.

Czaplewski, R. L. (1992), Accuracy assessment of remotely sensed classifications with multi-phase sampling and the multivariate composite estimator. In *Proceedings of the 16th International Biometrics Conference*, Hamilton, New Zealand, Volume 2, p. 22.

Czaplewski, R. L. (1994), Variance approximations for assessments of classification accuracy, Research Paper RM-316, USDA, Forest Service, Rocky Mountain Forest and Range Experiment Station, Fort Collins, CO, 29 pp.

Czaplewski, R. L. (1998), Accuracy assessments using two-phase stratified random sampling, cluster plots and the multivariate composite estimator. In *Spatial Accuracy in Natural Resource Analysis*, (H. T. Mowrer and R. G. Congalton, Eds.), Ann Arbor Press, Chelsea, MI, in press.

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.

Felix, N. A., and Binney, D. L. (1989), Accuracy assessment of a Landsat-assisted vegetation map of the coastal plain of the Arctic National Wildlife Refuge. *Photogramm. Eng. Remote Sens.* 55:475–478.

Fenstermaker, L. K. (1991), A proposed approach for national to global scale error assessments. Proc. GIS/LIS '91 1:293–300.

Fiorella, M., and Ripple, W. J. (1993), Determining successional stage of temperate coniferous forests with Landsat satellite data. *Photogramm. Eng. Remote Sens.* 59:239–246.

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

Fitzpatrick-Lins, K. (1981), Comparison of sampling procedures and data analysis for a land-use and land-cover map. *Photogramm. Eng. Remote Sens.* 47:343–351.

Franklin, S. E., Peddle, D. R., Wilson, B. A., and Blodgett,

C. F. (1991), Pixel sampling of remotely sensed digital imagery. *Comput. Geosci.* 17:759–775.

Fung, T., and LeDrew, E. (1988), The determination of optimal threshold levels for change detection using various accuracy indices. *Photogramm. Eng. Remote Sens.* 54:1449–1454.

George, T. H. (1986), Aerial verification of polygonal resource maps: A low-cost approach to accuracy assessment. *Photogramm. Eng. Remote Sens.* 52:839–846.

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.

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

Hord, R. M., and Brooner, W. (1976), Land-use map accuracy criteria. *Photogramm. Eng. Remote Sens.* 42:671–677.

Janssen, L. L. F., and van der Wel, F. J. M. (1994), Accuracy assessment of satellite derived land-cover data: A review. *Photogramm. Eng. Remote Sens.* 60:419–426.

Knick, S. T., Rotenberry, J. T., and Zarriello, T. J. (1997), Supervised classification of Landsat Thematic Mapper imagery in a semi-arid rangeland by nonparametric discriminant analysis. *Photogramm. Eng. Remote Sens.* 63:79–86.

Lauver, C. L., and Whistler, J. L. (1993), A hierarchical classification of Landsat TM imagery to identify natural grassland areas and rare species habitat. *Photogramm. Eng. Remote Sens.* 59:627–634.

Lepkowski, J., and Bowles, J. (1996), Sampling error software for personal computers. *The Survey Statistician* 35:10–17. (Also available through the American Statistical Association Survey Research Methods Section homepage, http://www.stat.ncsu.edu/info/srms/srms.html).

Martin, L. R. G. (1989), Accuracy assessment of Landsat-based visual change detection methods applied to the rural-urban fringe. *Photogramm. Eng. Remote Sens.* 55:209–215.

Martin, L. R. G., and Howarth, P. J. (1989), Change-detection accuracy assessment using SPOT multispectral imagery of the rural-urban fringe. *Remote Sens. Environ.* 30:55–66.

McGwire, K. C., Friedl, M., and Estes, J. E. (1993), Spatial structure, sampling design and scale in remotely-sensed imagery of a California savanna woodland. *Int. J. Remote Sens.* 14:2137–2164.

McGwire, K. C., and Estes, J. E., and Star, J. L. (1996), A comparison of maximum likelihood-based supervised classification strategies. *Geocarto Int.* 11(2):3–13.

Merchant, J. W., Yang, L., and Yang, W. (1993), Validation of continental-scale land cover data bases developed from AVHRR data. In *Proceedings of the PECORA Conference*, 24–26 August, pp. 63–72.

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 and Francis, New York, pp. 159–176.

Riley, R. H., Phillips, D. L., Schuft, M. J., and Garcia, M. C. (1997), Resolution and error in measuring land-cover change: effects on estimating net carbon release from Mexican terrestrial ecosystems. *Int. J. Remote Sens.* 18:121–137.

San Miguel-Ayanz, J., and Biging, G. S. (1996), An iterative classification approach for mapping natural resources from satellite imagery. *Int. J. Remote Sens.* 17:957–981.

San Miguel-Ayanz, J., and Biging, G. S. (1997), Comparison of single-stage and multi-stage classification approaches for cover type mapping with TM and SPOT data. *Remote Sens. Environ.* 59:92–104.

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

Senseman, G. M., Bagley, C. F., and Tweddale, S. A. (1995), Accuracy assessment of the discrete classification of remotely-sensed digital data for land-cover mapping, USACERL Technical Report EN-95/04, U.S. Army Corps of Engineers, Champaign, IL.

Smith, T. M. F. (1990), Comment on History and development of the theoretical foundations of survey based estimation and analysis. *Survey Methodol.* 16:26–29.

Stehman, S. V. (1992), Comparison of systematic and random sampling for estimating the accuracy of maps generated from remotely sensed data. *Photogramm. Eng. Remote Sens.* 58:1343–1350.

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), Use of auxiliary data to improve the precision of estimators of thematic map accuracy. *Remote Sens. Environ.* 58:169–176.

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. (1997a), Estimating standard errors of accuracy assessment statistics under cluster sampling. *Remote Sens. Environ.* 60:258–269.

Stehman, S. V. (1997b), Selecting and interpreting measures of thematic classification accuracy, *Remote Sens. Environ.* 62: 77–89.

Stehman, S. V., and Czaplewski, R. L. (1997), Basic structures of a statistically rigorous thematic accuracy assessment. In *Proceedings of the 1997 ACSM/ASPRS Annual Convention*, ASPRS Technical Papers, Vol. 3, pp. 543–553.

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

Stenback, J. M., and Congalton, R. G. (1990), Using thematic mapper imagery to examine forest understory. *Photogramm. Eng. Remote Sens.* 56:1285–1290.

Steven, M. D. (1987), Ground truth: an underview. *Int. J. Remote Sens.* 8:1033–1038.

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

Todd, W. J., Gehring, D. G., and Harmon, J. F. (1980), Landsat wild land mapping accuracy. *Photogrammetric Engineering and Remote Sensing* 46:509–520.

Verbyla, D. L., and Hammond, T. O. (1995), Conservative bias in classification accuracy assessment due to pixel-by-pixel comparison of classified images with reference grids. *Int. J. Remote Sens.* 16:581–587.

Vujakovic, P. (1987), Monitoring extensive 'buffer zones' in Africa: an application of satellite imagery. *Biol. Conservation* 39:195–208.

Walsh, S. J., Lightfoot, D. R., and Butler, D. R. (1987), Recognition and assessment of error in geographic information systems. *Photogramm. Eng. Remote Sens.* 53:1423–1430.

Warren, S. D., Johnson, M. O., Goran, W. D., and Diersing, V. E. (1990), An automated, objective procedure for selecting representative field sample sites. *Photogramm. Eng. Remote Sens.* 56:333–335.

Wickware, G. M., and Howarth, P. J. (1981), Change detection in the Peace-Athabasca Delta using digital Landsat data. *Remote Sens. Environ.* 11:9–25.

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

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

Zhuang, X., Engel, B. A., Xiong, X., and Johannsen, C. J. (1995), Analysis of classification results of remotely sensed data and evaluation of classification algorithms. *Photogramm. Eng. Remote Sens.* 61:427–433.