

Designing a Multi-Objective, Multi-Support Accuracy Assessment of the 2001
National Land Cover Data (NLCD 2001) of the Conterminous United States

Running Title: Accuracy Assessment of NLCD 2001

A framework for designing accuracy assessments of large-area land-cover maps is developed and
applied to the 2001 National Land Cover Data.

Stephen V. Stehman, Department of Forestry, 320 Bray Hall, SUNY-ESF, Syracuse NY 13210 USA
(svstehma@syr.edu)

James D. Wickham, National Exposure Research Laboratory, U.S. Environmental Protection Agency
(E243-05), RTP, NC 27711 USA (wickham.james@epamail.epa.gov)

Timothy G. Wade, National Exposure Research Laboratory, U.S. Environmental Protection Agency
(E243-05), RTP, NC 27711 USA (wade.timothy@epamail.epa.gov)

Jonathan D. Smith, Geographic Analysis and Monitoring Program, 12201 Sunrise Valley Dr., 519 USGS
National Center, Reston, VA 20192 USA (jhsmith@usgs.gov)

Abstract

The database design and diverse application of NLCD 2001 pose significant challenges for accuracy assessment because numerous objectives are of interest, including accuracy of land cover, percent urban imperviousness, percent tree canopy, land-cover composition, and net change. A multi-support approach is needed because these objectives require spatial units of different sizes for reference data collection and analysis. Determining a sampling design that meets the full suite of desirable objectives for the NLCD 2001 accuracy assessment requires reconciling potentially conflicting design features that arise from targeting the different objectives. Multi-stage cluster sampling provides the general structure to achieve a multi-support assessment, and the flexibility to target different objectives at different stages of the design. We describe the implementation of two-stage cluster sampling for the initial phase of the NLCD 2001 assessment, and identify gaps in existing knowledge where research is needed to allow full implementation of a multi-objective, multi-support assessment.

Key words: multi-stage sampling, land-cover composition, landscape pattern, impervious surface, canopy cover, land-cover change

Introduction

The Multi-Resolution Land Characteristics Consortium (MRLC) has completed the 2001 edition of the National Land Cover Data, NLCD 2001, for the conterminous United States (Homer *et al.* 2007), and is finishing NLCD 2001 for Alaska, Hawaii and Puerto Rico over the coming year (www.mrlc.gov). NLCD 2001 is considerably different from NLCD 1992 (Vogelmann *et al.* 2001). Whereas NLCD 1992 was a land-cover dataset, NLCD 2001 is a land-cover database consisting of three main products mapped for every 30-meter cell in the U.S. (Homer *et al.* 2004). The products identify one of 16 classes of land cover (Table 1) convolved to a one-acre minimum mapping unit, percent urban imperviousness, and percent tree canopy. The impervious surface and canopy cover datasets provide per-pixel estimates from 0 to 100 percent. In this article, we describe the framework and development of the sampling design for assessing accuracy of NLCD 2001.

Typically, the planning considerations for designing an accuracy assessment focus on obtaining adequate precision for estimating overall and class-specific accuracy (Stehman and Czaplewski 1998). Such *a priori* evaluations start by considering possible sampling designs and applying well-known variance equations (e.g. Cochran 1977) to gauge the sample size and allocation to balance cost constraints with desired precision. Because of NLCD's historical context (NLCD 1992 and 2001), the database design of NLCD 2001, and the widespread use of NLCD 1992 (Table 2), there is a proliferation of potential accuracy assessment objectives that extend beyond estimation of overall and class-specific accuracy. Therefore, the starting point for planning the NLCD 2001 accuracy assessment focuses on qualitative consideration of the interplay between multiple objectives, sampling design, and cost. Planning decisions must necessarily recognize the priorities among objectives because, in reality, available budgets will not keep pace with fiscal requirements needed to meet all objectives.

The sampling designs historically implemented to assess accuracy of large-area maps are not adequate to achieve the desired multi-objective, multi-support assessment of NLCD 2001. Multistage cluster sampling provides the structure to generate data at multiple supports, and the flexibility to

engineer the design to target multiple objectives. We describe how this design is used to advantage in the NLCD 2001 accuracy assessment. To fully implement a multi-objective, multi-support assessment, practical methods to achieve some of the desirable objectives must be developed, and the precision and cost of the multitude of possible sampling design options must be evaluated. We identify several of the more pressing needs for research and methodological development that would benefit subsequent phases of the NLCD 2001 accuracy assessment.

NLCD 2001 Accuracy Assessment Objectives and Sampling Design Criteria

The NLCD 2001 accuracy assessment objectives (Table 3) constitute a nearly exhaustive list of information on product quality. Meeting all of these objectives would leave little unknown about the quality of the three NLCD 2001 database elements, the reliability of change estimated from NLCD 1992 and 2001, the effect of reference data error (i.e. error in the determination of the “ground condition”) on the accuracy results, and the reliability of derivative measurements used in models and assessment. Realistically, available resources will limit the number of objectives that can be met, so the objectives were prioritized with that in mind. Objectives 1, 2, and 3, were given the highest priority because, as a set, they document product quality of the constituent elements of NLCD 2001 (Objectives 1 and 2) as well as a common application of the land-cover data (Objective 3).

The analyses that would accompany the top four priority objectives fall into two classes. Land-cover accuracy (Objective 1) would be addressed by the traditional analyses based on an error matrix. Land cover composition, net change, urban imperviousness, and tree canopy cover can all be expressed in terms of percent of area. Consequently, the analyses addressing Objectives 2-4 compare map quantities to reference data derived quantities. Descriptive accuracy measures such as mean deviation, mean absolute deviation, root mean square error, and correlation appropriately summarize these comparisons.

The desirable sampling design criteria for the NLCD 2001 assessment are listed in Table 4. Criteria 1, 2, and 3 are the most important because accuracy results need to be supported by the foundation of inferential statistics (Criterion 1), and confidence in the estimates should be as high as

possible (Criterion 2) given available fiscal resources (Criterion 3). Criteria 4 and 5 scale the first three criteria for national implementation, and Criterion 6 is included specifically to address Objective 5.

These design criteria in combination with the prioritized objectives guide the process for choosing the sampling design for the NLCD 2001 accuracy assessment. No one design will likely be compatible with all objectives and criteria. For example, the sampling and analysis spatial unit for Objectives 1 and 2 is typically a pixel, whereas Objectives 3-6 are better addressed by a larger spatial unit. In addition, precisely estimating land-cover composition accuracy (Objective 3) may require greater sampling effort than what is needed for estimating the per-pixel accuracy measures defined for Objectives 1 and 2. Objectives 5 and 6 require reference information for adjacent pixels, thereby introducing an additional structure beyond what is required by a sampling design targeting Objectives 1-4.

General Approaches to Design and Analysis of Map Accuracy Assessment

During the planning process, four general approaches to conducting an accuracy assessment typically surface. We categorize these approaches as cross validation, existing data, non-probability sampling, and probability sampling. Planning considerations may include questions such as, can the assessment be based solely on cross validation, are adequate reference data available from existing sources thus eliminating the need to collect new data, or is a judgment sample of a few intensively studied validation sites adequate to assess map accuracy? Evaluating the strengths and limitations of these approaches for a particular application is an important part of the planning process.

Cross Validation

Cross validation accuracy estimates are derived by omitting random subsets of the reference data used to train the classifier and then applying the resulting classification to the omitted subsets. Because the same data are used for training and accuracy assessment, the expense of obtaining a separate

independent reference sample is avoided, and the accuracy results are immediately available because the cross validation protocol is built into the classification process.

The statistical validity of the cross validation estimates is dependent on the sampling protocol used to obtain the reference data. If a probability sampling design is implemented, then the cross validation accuracy estimates are nearly unbiased (Steele 2005). Typically, however, the reference data incorporated in cross validation do not originate from a probability sample. Instead, locations are selected because of convenient access, to represent exemplars of a particular land-cover class, or because of homogeneity of land cover, and cross validation accuracy estimates derived from these data would not represent the true accuracy of the map. Steele (2005) and Baraldi *et al.* (2005) suggest analyses that can reduce the bias of cross validation estimates in this circumstance. If substantial manual re-interpretation of the output of the automated classification algorithm occurs, or if the final classification is the result of a complex procedure involving many steps combining different algorithms, the utility of the cross validation estimates is compromised.

The NLCD 2001 mapping protocol requires training data to drive the decision tree algorithms, and not all of these training data originate from a probability sampling design. Further, the NLCD 2001 classification procedures (Homer *et al.* 2004, plate 2) combine different algorithms and some manual re-interpretation has occurred. Cross-validation accuracy estimates are provided with the NLCD 2001 metadata, but users are cautioned not to interpret these estimates as representative of final product accuracy (Homer *et al.* 2007).

Existing Data

“Existing data,” defined as data initially obtained for some purpose other than accuracy assessment, have the potential advantage to lower the cost of reference data. Examples of accuracy assessments based on existing data include Balzter *et al.* (2002), Franklin *et al.* (2001), Fuller *et al.* (1998), Gill *et al.* (2000), and Wulder *et al.* (2006). Stehman *et al.* (2000) discuss conceptual and practical issues associated with using existing data. In the U.S., two high quality sources of existing data are the

U.S. Forest Service Forest Inventory and Analysis (FIA) (USFS 1992) and the National Resources Inventory (NRI) (Nusser and Goebel 1997). An appealing feature of both FIA and NRI is that they employ probability sampling designs. Other sources of existing data in the U.S. are less comprehensive than NRI and FIA, and often lack a probability sampling origin.

For a broad-based (i.e. national) accuracy assessment, existing data must be both compatible with the classification scheme of the map and sampled in a manner allowing the results to be generalized to the entire map area. Considerable effort may be needed to translate the FIA plot data and NRI classification of land cover into reference labels consistent with the NLCD 2001 classification scheme and spatial scale. Further, neither NRI nor FIA has complete national coverage. For example, the NRI does not apply to public lands, and therefore would not supply reference data for a significant portion of the NLCD 2001 map. Existing data are not entirely free because of the costs associated with evaluating suitability of the data, modifying the data for use, administratively managing data transfer, and filling gaps in the data (both spatially and by land-cover class). Existing data such as provided by FIA and NRI may contribute valuable information for some land-cover classes in some spatial domains. But creating a nationally consistent, complete accuracy assessment based exclusively on existing data would be a challenging task requiring patching together several existing data sources combined with supplemental sampling to fill in gaps in the data.

Non-probability Sampling

Judgment sampling and balanced sampling are non-probability sampling designs that may be considered for accuracy assessment. A judgment sample is purposely selected based on expert judgment to be representative of a population. Balanced sampling establishes representativeness quantitatively based on selecting the sample units so that characteristics of the sample (e.g., means and variances) approximately match those of the population.

Judgment sampling has been used to select validation sites employed for “confidence building measures” (i.e. studies or comparisons conducted to provide users with confidence in the map) (Morissette

et al. 2004). Validation sites may be chosen for convenience or to represent specific geographic areas or ecological conditions. A single validation site typically covers a large area (Muchoney *et al.* 1999), creating the potential to evaluate aspects of accuracy not generally available from the common practice of working with single pixels or small pixel blocks.

A large probability sample of validation sites would be ideal for accuracy assessment, but typically a limited number of validation sites are affordable. Cohen *et al.*'s (2006) examination of MODIS land cover accuracy at nine BIGFOOT sites is a good example of this feature of the validation site approach. Judgment sampling has merit when a very small sample size must be selected (Deming 1960), but inference must then be based either on a model or on the assumption that the sample is representative of the population, features which deter from a quantitatively rigorous assessment of accuracy. Although a few validation sites cannot provide a national, rigorous accuracy assessment, validation sites may contribute useful anecdotal information and play a valuable role in confidence-building measures.

Another variation of judgment sampling is to conduct the accuracy assessment within a limited geographic extent of the area mapped. For example, rather than conduct a national assessment for NLCD 2001, 5 of the 65 NLCD 2001 mapping zones could be selected to represent a diverse range of conditions encountered throughout the U.S. If a probability sampling design is implemented within each of five zones, then a rigorous basis exists for inference applicable to these five zones. The disadvantage is the limited geographical representation of the results. Nonetheless, if minimal resources are available to conduct a national assessment, a geographically restricted assessment based on a judgment sample of representative areas may be considered.

In general, balanced sampling seeks to achieve similarity between the sample statistics and corresponding population parameters of an auxiliary variable that is known for all elements of the population (Brewer 2002, p. 82, 102; Royall and Eberhardt 1975; Särndal *et al.* 1992, p. 531, 571). A simple approach is to balance on the mean of a single auxiliary variable, but more complex schemes exist to balance on several variables (Deville and Teillé 2004). Although randomization in the selection protocol is not necessary to support the model-based inferences associated with balanced sampling, it is

often advocated to lend objectivity (Valliant *et al.* 2000). Cihlar *et al.* (2000ab), Latifovic and Olthof (2004), and Latifovic *et al.* (2004) employed balanced sampling in which locations were selected so that the land-cover composition and a fragmentation index computed from the sample blocks matched the corresponding characteristics of the full region. Baret *et al.* (2006) used a balanced sampling approach to complete an existing network of sites for validating land biophysical products, selecting additional locations to balance the sample in terms of latitude and longitude, surface type (bare, grass, evergreen broadleaf, evergreen needleleaf, and deciduous broadleaf), and distribution of leaf area index.

Practical experience applying balanced sampling to accuracy assessment is limited. For example, what variables should be used to define balance and what is the rationale for choosing these variables? What model-based analyses are warranted, and what assumptions are inherent in these models? Design-based inference, a standard approach to analysis of accuracy assessment data (Stehman 2000), is predicated on an underlying probability sampling design, so implementing balanced sampling forfeits the option to apply these conventional analyses. Because model-based analyses traditionally have not been implemented in accuracy assessment and users are generally unfamiliar with these analyses, the NLCD 2001 sampling design should be chosen so that the assessment is not dependent on model-based analyses, yet does not preclude such analyses.

Probability Sampling

The probability sampling approach employs a rigorous protocol to select a sample of reference data. The protocol must satisfy two requirements: the inclusion probability of each element of the sample must be known, and the inclusion probability for any element of the population must be greater than 0, where an inclusion probability is the probability that a particular element of the population is included in the sample (Särndal *et al.* 1992, Stehman 2000). Simple random, stratified random, systematic, and cluster sampling are familiar probability sampling designs employed in accuracy assessment (Stehman 1999), and Table 5 lists examples of more complex probability sampling designs that have been used. Several design features are common to the Table 5 examples. Often the map land-cover classes are

defined as strata to target the objective of estimating user's accuracy, and cluster sampling is implemented to reduce costs.

A compelling advantage of the probability sampling approach is that it provides a strong statistical foundation for the assessment. Moreover, whereas design-based inference cannot be applied to a non-probability sample, data collected via a probability sampling design can be analyzed by model-based methodology. If model-based inference were to become more commonplace in accuracy assessment, the probability sampling approach would not prevent such analyses. The expense of collecting a sample of reference data independent from the classifier training data is the primary disadvantage.

The NLCD 1992 accuracy assessment (Stehman *et al.* 2003, Stehman 2004, Wickham *et al.* 2004a, Zhu *et al.* 2000) is an example of the probability sampling approach. We address the question of whether the NLCD 1992 design would meet the objectives and satisfy the desirable design criteria specified for the NLCD 2001 assessment. The NLCD 1992 design satisfied most of the desirable design criteria, starting with Criterion 1 (probability sampling design). The design was stratified by Level II land-cover class within each of 10 U.S. Environmental Protection Agency (EPA) Federal regions. For the Level II land-cover classification, the standard error for estimated overall accuracy (expressed as a proportion) was about 0.02 in each of the 10 regions, and the median standard error was 0.06 for all 180 user's accuracy estimates (10 EPA regions with an average of 18 classes per region). The standard errors for overall and user's accuracy for the Level I land-cover classes were generally smaller than the Level II standard errors. The ability to generate precise estimates of per class accuracy, even for classes that were rare in a region (Criterion 2), was a strength of the NLCD 1992 assessment.

Clustering the sample pixels within primary sampling units (PSUs) reduced costs (Criterion 3) by limiting the number of aerial photographs that had to be purchased and processed to obtain the reference land-cover information (Wickham *et al.* 2004b). The design was generally consistent nationally (Criterion 4) with minor deviations in protocol resulting from the sequential implementation of the design by the 10 EPA regions. A spatially well-distributed first-stage sample of PSUs across the U.S. (Criterion

5) resulted from stratifying by both the ten EPA regions and by geographic strata constructed within each region. The NLCD 1992 design was not constructed to collect reference data for blocks of adjacent pixels. Consequently, those analyses that require such data to evaluate the impact of reference data error (e.g., Hagen 2003) would not be available (Criterion 6).

The NLCD 1992 accuracy assessment was designed to meet Objective 1, and it successfully produced precise estimates of overall and class-specific accuracy for each of the 10 EPA Federal regions. Similar to other accuracy assessments (Table 5), the NLCD 1992 design was not constructed to meet the expanded set of objectives proposed for NLCD 2001. However, the NLCD 1992 design serves as a good starting point for a multi-support, multi-objective design. Not only does it satisfy most of the desirable design criteria, the NLCD 1992 design provides a suitable structure that, with slight modification, can accommodate the additional objectives of NLCD 2001.

Sampling Designs for Multi-Support, Multi-Objective Accuracy Assessment

Multi-stage cluster sampling (Cochran 1977, Chapter 10) provides a structure compatible with the objectives of a multi-support accuracy assessment. To simplify discussion of the sampling design issues, we restrict attention to two levels of support, the 30 m pixel and one larger areal support unit. In general, the larger support unit could be a county, a watershed, or a regular tessellation unit such as a square or hexagon. In cluster sampling terminology, the larger support unit is the primary sampling unit (PSU), and the pixel is the secondary sampling unit (SSU). In a one-stage cluster sampling design for an assessment targeting two levels of support, all pixels within the first-stage sample PSUs are included in the sample, and reference land-cover data are obtained for the full PSU. This design makes the assessment of composition accuracy and net change accuracy straightforward because the reference classification of composition and net change of each PSU is known rather than estimated from a subsample of the PSU, as would be necessary in two-stage cluster sampling.

The major deterrent to employing one-stage cluster sampling is the size of the PSU that could be used in practice. Because all pixels within each sampled PSU must be interpreted, the PSU size must necessarily be small. If the PSU size is large, each PSU becomes costly to interpret, and too few sample PSUs may be affordable detracting from the precision of the accuracy estimates. The difficulty with a small PSU is that it may not represent a practically useful support for the composition and net change objectives. For example, a 10 by 10 pixel PSU would be manageable from an interpretation standpoint, but would represent a support of only 0.09 km² (0.3 km by 0.3 km). A more practically relevant support for applications of land-cover composition would be 9 km² (3 km by 3 km), requiring a 100 by 100 pixel PSU. The small PSU size required for one-stage cluster sampling to be practical motivates consideration of two-stage cluster sampling. Another disadvantage of the one-stage design is that cluster sampling generally increases the standard errors of accuracy estimators unless the reduced per-pixel cost of cluster sampling compensates for the precision loss attributable to within cluster correlation of the response (Moisen *et al.* 1994). Two-stage cluster sampling can diminish the variance inflation feature of one-stage cluster sampling.

Numerous two-stage sampling designs combining different first- and second-stage designs are available, and a variety of design criteria may affect the choice of design for each stage. The first-stage design will be an equal probability sample of PSUs selected from a population of equal-sized PSUs that partition the region of interest. Two options for the second-stage design are constructed to target one of two objectives, class-specific accuracy (Option A) or composition accuracy (Option B). In design Option A, the pixels within the first-stage sample of PSUs are stratified by Level II land-cover class. A stratified random sample of pixels is selected, with the sample size for each Level II class specified (e.g., 100 pixels per class). To target the composition accuracy objective, the sample size selected per PSU in Option B is specified (e.g., 50 pixels per PSU), but stratification by land-cover class is not employed. Option B could be implemented by selecting a simple random sample of pixels from each sampled PSU.

The stratification employed in Option A controls the sample size allocated per land-cover class, but does not control the number of pixels sampled from each PSU. Depending on the spatial distribution

and rarity of a land-cover class, some PSUs will have very few sample pixels, while others will receive a large number. Option A may perform poorly for Objective 3 because too few sample pixels will be available in some PSUs to estimate land-cover composition. Because the number of pixels sampled from each PSU is specified, Option B permits stronger control for the composition accuracy objective. But because Option B is an equal probability sampling design, the sample size realized for any given land-cover class will be proportional to that class's representation in the population. Unless the overall sample size is very large, an unlikely situation given the limited budget for accuracy assessment, the sample sizes for rare classes will be small and the accuracy estimates for these classes (Objective 1) will be imprecise.

Accuracy assessments (Table 5) have not been designed to achieve both Objectives 1 and 3 simultaneously, and Objective 3 has been attempted only in situations where complete coverage reference data were available (Hollister *et al.* 2004, Johnson *et al.* 1999). Further investigation is needed to determine what designs and allocation rules best simultaneously address Objectives 1 and 3. A design combining features of both Options A and B may be necessary to ensure that a specified sample size is obtained per sampled PSU and per land-cover class. For example, if Option B is implemented, those land-cover classes that have received only a few of the second-stage sample pixels could be targeted for supplemental sampling, applying the stratified sampling approach of Option A. If Option A is implemented, those PSUs not receiving an adequate sample size to estimate land-cover composition could be targeted for supplemental sampling by selecting additional sample pixels from those PSUs to increase the sample size to the desired target. Simplicity is a desirable design criterion because resource and time limitations prohibit extensive investigation of alternative design options and development of complex analysis protocols. Simple designs are more likely to be implemented properly and analyzed correctly, and they are better able to accommodate changed objectives and uncertain budgets.

NLCD 2001 Accuracy Assessment Implementation

A two-stage cluster sampling design using Option A will be implemented for the NLCD 2001 assessment. Cluster sampling was chosen to reduce costs by spatially constraining the sample to a fixed number of PSUs, and also to provide data at a larger support (the PSU size) to allow for estimating composition accuracy. Although the primary source of reference data, 1-meter Digital Orthoimagery Quarter Quadrangles (DOQQs), will often be accessible online, cluster sampling is still expected to reduce the time required to acquire and process reference data because the sample pixels will be grouped within fewer DOQQs.

The first-stage sampling design is a geographically stratified random sample of PSUs, chosen to spatially balance the sample across the U.S. (Figure 1). The PSU is a 24 km by 24 km block, chosen for reasons discussed later in this section. Each geographic stratum is 240 km by 120 km, and one PSU is selected at random with equal probability from the 50 PSUs available in each stratum. All strata are constructed to contain a full complement of 50 PSUs, so a stratum may extend beyond the U.S. border (i.e., non-target PSUs are included in the sampling frame). If the randomly selected PSU for a stratum falls outside the U.S., no sample PSU is obtained for that stratum. This maintains the equal inclusion probability property for all PSUs in the U.S. The ten accuracy assessment regions (Figure 1) represent another layer of geographic stratification relevant to the second-stage of the design. These ten regions were delineated on the basis of similar land-cover composition and pattern, and each region has nested within it a subset of the NLCD 2001 mapping zones.

Because the primary objective for the NLCD 2001 assessment is per-class accuracy of land cover, the sampling design is stratified by Level II land-cover class within each Figure 1 accuracy assessment region. Stratifying by both region and Level II land-cover class will allow for precise estimates of accuracy of the Level II classes even in those regions where a class is not abundant (Criteria 2, 4, and 5). Two examples illustrate the rationale for this decision. Deciduous forest (class 41) is common in Region 8 (49.5% of the area) but rare in Region 3 (3.42% of the area). Nationally, 33% of the class 41 pixels in the first-stage sample are in Region 8, and 2.5% are in Region 3. If the second-stage sample for deciduous forest is selected with equal probability from the deciduous forest pixels available nationally

(assuming 1000 deciduous forest pixels are sampled nationally), approximately 330 of the deciduous forest sample pixels (33%) will be from Region 8, and only 25 sample pixels (2.5%) will be from Region 3. Cropland (class 82) is dominant in Region 5 (51.1% of the area), but relatively rare in Region 3 (0.66% of the area). The second-stage sample for cropland, if not stratified by accuracy assessment region, would result in about 50% of the sample pixels from Region 5 and 0.7% (7 pixels) from Region 3.

The 24 km by 24 km PSU was chosen based on considerations of both Objectives 1 and 3 and cost. Estimating accuracy of land-cover composition (Objective 3) requires a support size representative of the applications in which NLCD 2001 land-cover information is aggregated. Typical applications use support sizes ranging from 5 km by 5 km to 50 km by 50 km (the mean size of a county in the U. S.). The 24 km by 24 km PSU falls toward the lower end of this range of support sizes. Based on our initial evaluation of precision for estimating land-cover composition accuracy, a sample size of about 50 pixels per PSU is needed. Thus the sampling design should generate, on average, 50 sample pixels per PSU. Given an anticipated budget allowing for 1,500 sample pixels in each of the 10 accuracy assessment regions, approximately 30 PSUs will be sampled per region to meet Objective 3. To better satisfy Objective 1 (per-class accuracy of land cover), it is desirable that the sample pixels be spatially dispersed to reduce the impact of within-PSU correlation on precision. A larger PSU creates a larger pool of pixels from which to select the second-stage sample and provides a greater spatial dispersion of the sample pixels. The 24 km by 24 km PSU was deemed sufficiently large for this purpose. The decision to use a 24 km by 24 km PSU represents a compromise choice: It is small enough that some cost advantage accrues from grouping the second-stage sample within fewer DOQQs, and it is large enough that the second-stage sample pixels are more spatially dispersed than would be the case for, say, a 6 km by 6 km PSU.

Accuracy of percent urban imperviousness and percent tree canopy (Objective 2)

The sampling design described in the previous section can be used to assess accuracy of the urban imperviousness and tree canopy products. The per-pixel cost to obtain the reference data for these

quantitative databases will be higher than the per-pixel cost for land cover because percent urban imperviousness and percent tree canopy will require more complicated and time-consuming reference data protocols. Consequently, the sample size for this objective must be reduced to maintain affordability.

Separate subsamples selected from the base sample chosen for assessing land-cover accuracy will be used to assess the imperviousness and canopy products. Each subsample will be stratified to ensure that the subsample covers the full range of values of percent urban imperviousness and percent tree canopy. The strata for the impervious surface subsample will be the Level II land-cover classes 21, 22, 23, and 24 (Table 1). These classes are defined by percent impervious values of 0-20% (Class 21), 21-50% (Class 22), 50-80% (Class 23), and 80-100% (Class 24). A fifth stratum will be comprised of all the non-urban land-cover classes. The stratified subsample will be equally allocated to these five strata. The strata for the canopy density subsample will be 0-40%, 41-60%, 61-80%, 81-90%, and 91-100%, representing approximately equal proportions of the U.S. in each stratum (Homer *et al.* 2007, Table 1). The subsample will be allocated equally to these strata. The ten accuracy assessment regional strata will also apply to the subsampling design to improve estimates of the accuracy of urban imperviousness and tree canopy classifications that may be regionally uncommon.

Accommodating Objective 2 raises no new sampling considerations for choosing the base sampling design for NLCD 2001. The only modification required is the additional subsampling from the two-stage cluster sample selected to assess land-cover accuracy. The total sample size allocated to assessing accuracy of the imperviousness and canopy products will depend on the cost to obtain the reference data and the available budget.

Accuracy of Net Change (Objective 4)

The two-stage cluster sampling design can also be used to assess net change accuracy for the Level I classes at the support of the 24 km by 24 km PSU. The net change assessment requires obtaining the 1992 Level I reference land-cover class for each sampled pixel. Similar to the assessment of land-

cover composition accuracy, the net change assessment requires that each sampled PSU receives an adequate sample size of pixels to estimate net change, 50 being the established size for estimating composition accuracy. Obtaining reference data for both 1992 and 2001 increases the cost of the assessment considerably, but reducing costs by decreasing the sample size is untenable because the number of PSUs sampled per region (30) and the number of SSUs sampled per PSU (50) is likely already at the lower end of what is needed to obtain adequate precision.

Because change is typically rare, stratifying the PSUs by map net change would allow for increasing the number of high net change PSUs sampled. The two-stage cluster sampling design offers the flexibility to stratify the PSUs to target the net change accuracy objective, while still stratifying pixels within PSUs to target the class-specific accuracy objective. The NLCD 2001 design is not stratified by net change, but we are planning to explore the effect stratifying PSUs by net change has on precision of the estimates for land-cover class accuracy (per pixel) and composition accuracy (per PSU).

Accuracy of Landscape Features and Pattern (Objective 6)

The now well-established use of derivative map measurements (e.g., land-cover patch size) in models and national environmental assessments (Heinz Center 2002, USEPA 2003) has motivated research into how to assess their accuracy. The most basic attribute of pattern is edge – the boundary between two pixels with dissimilar attribution (e.g., agriculture adjacent to urban). Edge accuracy is the primary information needed to estimate accuracy of the numerous pattern measurements (Hess 1994, Hess and Bay 1997, Wickham *et al.* 1997) calculated in off-the-shelf software packages (e.g., FRAGSTATS, McGarigal and Marks 1995). More broadly, any technique for comparing two maps (i.e. quantifying their similarity) could potentially be used to quantify accuracy, where the reference data are considered the map to compare with the target map. These map comparison techniques include methods to quantify agreement of feature attribute, location, size, and shape, and relationships among features (Dungan 2006, Hargrove *et al.* 2006), agreement of the land-cover proportions at multiple scales (Pontius

and Cheuk 2006), and agreement of patterns based on areal intersection of land use polygons (White 2006) or fuzzy local matching of polygons (Power *et al.* 2001).

Because landscape pattern and feature based accuracy assessments have not been conducted for large-area land-cover maps, there is little research or experience to suggest what size assessment unit is appropriate. The choice of assessment unit will have a strong impact because truncation of land-cover polygons and other edge effects introduced by the unit boundary will affect the values of the pattern and feature metrics and accordingly the accuracy measures derived for these metrics. Practical experience is also lacking when it comes to obtaining reference data for large pixel blocks in a national level assessment such as NLCD 2001. It is certainly feasible to obtain quality reference data for small blocks of adjacent pixels (e.g. 3 by 3 or 5 by 5), but what practical concerns arise if a relatively large block, say, 25 by 25 pixels, must be interpreted? Similar considerations extend to a sampling design tailored to Objective 5. For example, methods suggested for accommodating reference data error in the analysis require reference data for a block of pixels (Hagen 2003, Hagen-Zanker 2006).

One-stage cluster sampling would merit strong consideration if Objectives 5 and 6 are considered high priorities. Because all pixels within a PSU are sampled in this design, it is possible to directly compute features such as patch size, shape, and edge, as well as relationships among these features. In contrast, it may be difficult to accurately derive characteristics and relationships among features from a two-stage sample of spatially separate pixels. The concern remains that the PSU chosen for one-stage cluster sampling has to be large enough to be of practical use for the composition and net change accuracy objectives. A solution to this problem would be to design the assessment to address three levels of support, the smallest being a pixel, the largest being a PSU concordant with the applications requiring composition and net change accuracy, and an intermediate size to address Objectives 5 and 6 (all pixels in this intermediate size unit would be sampled). However, targeting three levels of support would further exacerbate the design trade-offs among objectives, and additional work is needed to establish if a three-level support assessment is feasible.

Phasing in the NLCD 2001 Accuracy Assessment

Large-area mapping projects typically are not adequately funded to implement a full accuracy assessment (Scepan 1999). Instead, accuracy assessment is tacked on at the end of the mapping process, with the amount and timing of funding uncertain. Consequently, it is often necessary to implement the assessment sequentially over several years, with the uncertainties related to future funding taken into account when deciding how to phase in the design. Initial funding for NLCD 2001 accuracy assessment will support documenting the base level descriptive accuracy of NLCD 2001 in Region 8 (Figure 1) focusing on Objectives 1 and 3. Region 8 becomes an operational field test of the sampling design and reference data collection protocols of a multi-support accuracy assessment on the way to implementing a full national assessment. Results from Region 8 will indicate whether the sampling design needs to be adjusted or augmented to better address Objectives 1 and 3. The sampling design chosen for the NLCD 2001 assessment is flexible to incorporate modifications if the net change objective becomes relevant. Although a subsampling design is in place to address accuracy of the urban imperviousness and tree canopy products, initial funding will not support evaluating these products in the first phase of the assessment. The Region 8 sample size is 1500 pixels (100 per Level II class), commensurate with the sampling effort expended per EPA Federal region in the assessment of NLCD 1992.

Conclusions

A sampling design that accommodates multiple accuracy assessment objectives at different levels of support is needed to fully evaluate NLCD 2001. Current practice for assessing accuracy of large-area maps achieves the objective of estimating per-class and overall accuracy, but assessments have typically not been designed to simultaneously address an expanded set of objectives that require both pixel level and larger areal supports. The two-stage cluster sampling design has the flexibility to tailor the first-stage design to target the larger support (PSU) objectives of net change and composition accuracy, and the second-stage design can be targeted for the pixel level (SSU) support objectives evaluating accuracy of

land cover, percent urban imperviousness, and percent tree canopy. Achieving a balance among the potentially conflicting sampling design requirements of these objectives is a critical design consideration.

Planning for the NLCD 2001 design took into account variance and cost issues qualitatively. But several obstacles hinder conducting a meaningful quantitative comparison of design options. The target parameters for comparing design alternatives are overall accuracy and user's accuracy for land cover, and root mean square error or mean absolute deviation for percent urban imperviousness, percent tree canopy, land-cover composition, and net change. Variance approximations for the sample-based estimators of land-cover accuracy have been derived for designs akin to the two-stage cluster sampling design proposed for the NLCD 2001 assessment (Nusser and Klaas 2003, section 3.4.4; Stehman *et al.* 2003, Appendix A). These variances are complicated for the estimators of overall accuracy and user's accuracy. The variance expressions for estimated root mean square error or mean absolute deviation have not been derived and will be complicated for multi-stage sampling. Consequently, analytic comparison of design alternatives based on the variances of the estimated target parameters is not a realistic option to guide the selection of the sampling design for NLCD 2001. A simulation approach comparing variances under different design options is a practical alternative, but even this will be extremely challenging because it requires creating or identifying populations that have magnitudes and spatial patterns of classification error representative of conditions found in products such as the NLCD 2001 databases.

Although multi-stage cluster sampling provides a general design structure to accommodate the data needs of a multi-objective, multi-support assessment, several methodological developments must occur before such an assessment can be fully implemented. These include procedures for estimating accuracy of land-cover composition and net change for the multi-stage sampling design, and practical methods for obtaining sample-based estimators of accuracy of landscape features and pattern. It is also necessary to evaluate quantitatively design alternatives that achieve the multiple accuracy objectives in a cost efficient, integrated manner. Establishing a general design framework creates the present ability to implement a multi-support, multi-objective accuracy assessment. Identifying the research issues required

to fully achieve all objectives creates the opportunity to improve upon the methods implemented in the initial phase of the NLCD 2001 assessment.

Acknowledgments

We thank the four anonymous reviewers for their frank and useful comments on the manuscript. Surely we have not satisfied all of their concerns, but we greatly appreciate the effort and thought invested in their suggestions. This article has not been subject to EPA or USGS review and does not necessarily reflect the views of either agency.

References

- Alexander, R. B., Smith, R. A., and Schwarz, G. E., 2004. Estimates of diffuse phosphorus sources in surface waters of the United States using a spatially referenced watershed model, *Water Science and Technology*, 49:1-10.
- Amundson, R., Guo, Y., and Gong, P., 2003. Soil diversity and land use in the United States, *Ecosystems*, 6:470-482.
- Balzter, H., Talmon, E., Wagner, W., Gaveau, D., Plummer, S., Yu, J. J., Quegan, S., Davidson, M., Toan, T. L., Gluck, M., Shvidenko, A., Nilsson, S., Tansey, K., Luckman, A., and Schmullius, C., 2002. Accuracy assessment of a large-scale forest cover map of central Siberia from synthetic aperture radar, *Canadian Journal of Remote Sensing*, 28:719-737.
- Baraldi, A., Bruzzone, L., and Blonda, P., 2005. Quality assessment of classification and cluster maps without ground truth knowledge, *IEEE Transactions on Geoscience and Remote Sensing*, 43:857-873.
- Baret, F., Morisette, J. T., Fernandes, R. A., Champeaux, J. L., Myneni, R. B., Chen, J., Plummer, S., Weiss, M., Bacour, C., Garrigues, S., and Nickeson, J. E., 2006. Evaluation of the representativeness of networks of sites for the global validation and intercomparison of land

- biophysical products: Proposition of the CEOS-BELMANIP, *IEEE Transactions on Geoscience and Remote Sensing*, 44:1794-1803.
- Biging, G. S., Colby, D. R., and Congalton, R. G., 1998. Sampling systems for change detection accuracy assessment, *Remote Sensing Change Detection: Environmental Monitoring Methods and Applications* (R. S. Lunetta and C. D. Elvidge, editors), Ann Arbor Press, Chelsea, Michigan, pp. 281-308.
- Brabec, E., Schulte, S., and Richards, P., 2002. Impervious surfaces and water quality: a review of current literature and its implications for watershed planning, *Journal of Planning Literature*, 16:499-514.
- Brewer, K. 2002. *Combined Survey Sampling Inference: Weighting Basu's Elephants*, Arnold Publishers, London, 316 pp.
- Cihlar, J., Latifovic, R., Chen, J., Beaubien, J., and Li, Z., 2000a. Selecting representative high resolution sample images for land cover studies. Part 1: Methodology, *Remote Sensing of Environment*, 71:26-42.
- Cihlar, J., Latifovic, R., Chen, J., Beaubien, J., Li, Z., and Magnussen, S., 2000b. Selecting representative high resolution sample images for land cover studies. Part 2: Application to estimating land cover composition, *Remote Sensing of Environment*, 72:127-138.
- Cochran, W. G., 1977. *Sampling Techniques* (3rd ed.), John Wiley & Sons, New York.
- Cohen, W. B., Maersperger, T. K., Turner, D. P., Ritts, W. D., Pflugmacher, D., Kennedy, R. E., Kirschbaum, A., Running, S. W., Costa, M., and Gower, S. T., 2006. MODIS land cover and LAI Collection 4 product quality across nine sites in the western hemisphere, *IEEE Transactions on Geoscience and Remote Sensing*, 44:1843-1857.
- Congalton, R. G., and Green, K., 1993. A practical look at the sources of confusion in error matrix generation. *Photogrammetric Engineering & Remote Sensing*, 59:641-644.
- DeLiberty, T. L., and Legates, D. R., 2003. Interannual and seasonal variability of modelled soil moisture in Oklahoma, *International Journal of Climatology*, 23:1057-1086.

- Deming, W. E., 1960, *Sampling Design in Business Research*, John Wiley & Sons, New York.
- Deville, J.-C., and Tillé, Y., 2004. Efficient balanced sampling: The cube method, *Biometrika*, 91:893-912.
- Dungan, J. L., 2006. Focusing on feature-based differences in map comparison, *Journal of Geographical Systems*, 8:131-143.
- Durbrow, R. P., Burns, N. B., Richardson, J. R., and Berish, C. W., 2001. Southeastern ecological framework: a planning tool for managing ecosystem integrity. Proceedings of the 2001 Georgia Water Resources Conference, March 26-27, 2001, University of Georgia, Athens, GA.
- Edwards, T. C., Jr., Moisen, G. G., and Cutler, D. R., 1998. Assessing map accuracy in an ecoregion-scale cover-map, *Remote Sensing of Environment*, 63:73-83.
- Franklin, J., Beardsley, D., Gordon, H., Simons, D. K., and Rogan, J. M., 2001. Evaluating errors in a digital vegetation map with forest inventory data and accuracy assessment using fuzzy sets, *Transactions in GIS*, 5:285-304.
- Fuller, R. M., Wyatt, B. K., and Barr, C. J., 1998. Countryside survey from ground and space: different perspectives, complementary results, *Journal of Environmental Management*, 54:101-126.
- Gill, S. J., Milliken, J., Beardsley, D., and Warbington, R., 2000. Using a mensuration approach with FIA vegetation plot data to assess the accuracy of tree size and crown closure classes in a vegetation map of Northeastern California, *Remote Sensing of Environment*, 73:298-306.
- Hagen, A., 2003. Fuzzy set approach to assessing similarity of categorical maps, *International Journal of Geographical Information Science*, 17:235-249.
- Hagen-Zanker, A., 2006. Map comparison methods that simultaneously address overlap and structure, *Journal of Geographical Systems*, 8:165-185.
- Hargrove, W. W., Hoffman, F. M., and Hessburg, P. F., 2006. Mapcurves: a quantitative method for comparing categorical maps, *Journal of Geographical Systems*, 8:187-208.
- Heilman, G. E., Jr., Strittholt, J. R., Slosser, N.C., and Dellasala, D.A. 2002. Forest fragmentation of the

- conterminous United States: assessing forest intactness through road density and spatial characteristics, *BioScience*, 52:411-422.
- Heinz Center (The H. John Heinz III Center for Science, Economics, and the Environment), 2002. *The State of the Nation's Ecosystems: Measuring the Lands, Waters and Living Resources of the United States*, Cambridge University Press, Cambridge, UK.
- Hess, G. R., 1994. Pattern and error in landscape ecology: A commentary, *Landscape Ecology*, 9:305.
- Hess, G. R., and Bay, J. M., 1997. Generating confidence intervals for some measures of landscape pattern, *Landscape Ecology*, 12:309-320.
- Hollister, J. W., Gonzalez, M. L., Paul, J. F., August, P. V., and Copeland, J. L., 2004. Assessing the accuracy of National Land Cover Dataset area estimates at multiple spatial extents, *Photogrammetric Engineering & Remote Sensing*, 70: 405-414.
- Homer, C., Huang, C., Yang, L., Wylie, B., and Coan, M., 2004. Development of a 2001 National Land-Cover Database for the United States, *Photogrammetric Engineering & Remote Sensing*, 70:829-840.
- Homer, C., Dewitz, J., Fry, J., Coan, M., Hossain, N., Larson, C., Herold, N., McKerrow, A., VanDriel, J. N., and Wickham, J. D., 2007. Completion of the 2001 National Land Cover Database for the conterminous United States, *Photogrammetric Engineering & Remote Sensing*, 73: 337-341.
- Johnson, C. B., Weaver, G. W., and Larsen, D. P., 1999. A comparison of land use and land cover data in watersheds of the Mid-Atlantic region, USA, *Environmental Practice*, 1: 209-226.
- Latifovic, R., and Olthof, I., 2004. Accuracy assessment using sub-pixel fractional error matrices of global land cover products derived from satellite data, *Remote Sensing of Environment*, 90:153-165.
- Latifovic, R., Zhu, Z., Cihlar, J., Giri, C., and Olthof, I., 2004. Land cover mapping of North and Central America - Global Land Cover 2000. *Remote Sensing of Environment*, 89:116-127.
- Lawler, J. J., White, D., and Master, L. L., 2003. Integrating representation and vulnerability: two approaches for prioritizing areas for conservation, *Ecological Applications*, 13:1762-1772.

- Marshall, C. H., Pielke, Sr., R. A., Steyaert, L. T., and Willard, D. A., 2004. The impact of anthropogenic land-cover change on the Florida peninsula sea breezes and warm season sensible weather, *Monthly Weather Review*, 132:28-52.
- Mayaux, P., Eva, H., Gallego, J., Strahler, A. H., Herold, M., Agrawal, S., Naumov, S., De Miranda, E. E., Di Bella, C. M., Ordoyne, C., Kopin, Y., and Roy, P. S., 2006. Validation of the Global Land Cover 2000 map, *IEEE Transactions on Geoscience and Remote Sensing*, 44:1728-1739.
- Mayer, A. L. and Cameron, G. N., 2003. Landscape characteristics, spatial extent, and breeding bird diversity in Ohio, USA, *Diversity and Distributions* 9:297-312.
- McGarigal, K., and Marks, B. J., 1995. FRAGSTATS: spatial pattern analysis program for quantifying landscape structure. USDA Forest Service General Technical Report PNW-351.
- Moisen, G. G., Edwards, Jr., T. C., 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.
- Morisette, J. T., Privette, J. L., Strahler, A., Mayaux, P., and Justice, C. O., 2004. Validation of global land-cover products by the Committee on Earth Observing Satellites, in *Remote Sensing and GIS Accuracy Assessment*, R. S. Lunetta and J. G. Lyon (eds.), CRC Press, New York, pp. 31-40.
- Muchoney, D. M., Strahler, A. H., Hodges, J., and LoCastro, J., 1999. The IGBP DISCover confidence sites and the system for terrestrial ecosystem parametrization: Tools for validating global land-cover data, *Photogrammetric Engineering & Remote Sensing*, 65:1061-1067.
- Nolan, B. T., Hitt, K. J., and Ruddy, B. C., 2002. Probability of nitrate contamination of recently recharged ground waters in the conterminous United States, *Environmental Science and Technology*, 36:2138-2145.
- Nusser, S. M., and Goebel, J. J., 1997. The National Resources Inventory: a long-term multi-resource monitoring program, *Environmental and Ecological Statistics*, 4:181-204.
- Nusser, S. M., and Klaas, E. E., 2003. Survey methods for assessing land cover map accuracy,

- Environmental and Ecological Statistics*, 10:309-331.
- Pontius, R. G., Jr., and Cheuk, M. L., 2006. A generalized cross-tabulation matrix to compare soft-classified maps at multiple spatial resolutions, *International Journal of Geographical Information Science*, 20:1-30.
- Powell, R. L., Matzke, N., de Souza, Jr., C., Clark, M., Numata, I., Hess, L. L., and Roberts, D. A., 2004. Sources of error in accuracy assessment of thematic land-cover maps in the Brazilian Amazon, *Remote Sensing of Environment*, 90:221-234.
- Power, C., Simms, A., and White, R., 2001. Hierarchical fuzzy pattern matching for the regional comparison of land use maps, *International Journal of Geographical Information Science*, 15:77-100.
- Radeloff, V. C., Hammer, R. B., Stewart, S. I., Fried, J. S., Holcomb, S. S., and McKeefry, J. F., 2005. The wildland-urban interface in the United States, *Ecological Applications*, 15:799-805.
- Risser, P. G., Karr, J. R., and Forman, R. T. T., 1984, Landscape ecology: Directions and approaches, Illinois Natural History Survey Special Publication Number 2, Champaign, IL.
- Riitters, K. H., Wickham, J. D., O'Neill, R. V., Jones, K. B., Smith, E. R., Coulston, J. W., Wade, T. G., and Smith, J. H., 2002. Fragmentation of continental United States forests, *Ecosystems*, 5: 815-822.
- 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.
- SC DHEC, 2004. Total Maximum Daily Load for Big Wateree Creek: Station CW-072 Fecal Coliform. South Carolina Department of Health and Environmental Control.
- Scepan, J., 1999. Thematic validation of high-resolution global land-cover data sets, *Photogrammetric Engineering & Remote Sensing*, 65:1051-1060.
- Schmidt, K. M., Manakis, J. P., Hardy, C. C., Hann, W. J., and Bunnell, D. L., 2002. Development of

- coarse-scale spatial data for wildland fire and fuel management. United States Department of Agriculture, Forest Service, Rocky Mountain Research Station, Gen. Tech. Report RMRS-87, Fort Collins, CO.
- Sheeder, S. A., Lynch, J. A., and Grimm, J., 2002. Modeling Atmospheric nitrogen deposition and transport in the Chesapeake Bay Watershed, *Journal of Environmental Quality*, 31:1194-1206.
- Smith, S. V., Renwick, W. H., Bartley, J. D., and Buddemeier, R. W., 2002. Distribution and significance of small, artificial water bodies across the United States landscape, *The Science of the Total Environment*, 299:21-36.
- Steele, B. M., 2005. Maximum posterior probability estimators of map accuracy, *Remote Sensing of Environment*, 99:254-270.
- Stehman, S. V., 1999. Basic probability sampling designs for thematic map accuracy assessment, *International Journal of Remote Sensing*, 20: 2423-2441.
- Stehman, S. V., 2000. Practical implications of design-based sampling inference for thematic map accuracy assessment, *Remote Sensing of Environment*, 72:35-45.
- Stehman, S. V., 2001. Statistical rigor and practical utility in thematic map accuracy assessment, *Photogrammetric Engineering & Remote Sensing*, 67:727-734.
- Stehman, S. V., 2004. Sampling designs for accuracy assessment of large-area, land-cover maps: Challenges and future directions, *Remote Sensing and GIS Accuracy Assessment* (R. S. Lunetta and J. G. Lyon, editors), CRC Press, New York, pp. 13-29.
- Stehman, S. V., and Czaplewski, R. L., 1998. Design and analysis for thematic map accuracy assessment: Fundamental principles, *Remote Sensing of Environment*, 64: 331-344.
- Stehman, S. V., Czaplewski, R. L., Nusser, S. M., Yang, L., and Zhu, Z., 2000. Combining accuracy assessment of land-cover maps with environmental monitoring programs, *Environmental Monitoring and Assessment*, 64:115-126.
- Stehman, S. V., and Wickham, J. D., 2006. Assessing accuracy of net change derived from land cover maps, *Photogrammetric Engineering & Remote Sensing*, 72:175-185.

- Stehman, S. V., Wickham, J. D. Smith, J. H., and Yang, L., 2003. Thematic accuracy of the 1992 National Land-Cover Data (NLCD) for the Eastern United States: Statistical methodology and regional results, *Remote Sensing of Environment*, 86:500-516.
- Treves, A., Naughton-Treves, L., Harpers, E. K., Mladenoff, D. J., Rose, R. A., Sickley, T. A., and Wydeven, A. P., 2004. Predicting human-carnivore conflict: a spatial model derived from 25 years of data on wolf predation on livestock, *Conservation Biology*, 18:114-125.
- USEPA, 2003. *EPA's Draft Report on the Environment: Technical Document*, EPA 600-R-03-050, U. S. Environmental Protection Agency, Washington, DC
(<http://www.epa.gov/indicators/publications.htm> , last viewed 1 June 2007).
- USFS, 1992. *Forest Service Resource Inventories: An Overview*, USGPO 1992-341-350/60861, U. S. Department of Agriculture, Forest Service, Forest Inventory, Economics, and Recreation Research, Washington, DC.
- Valliant, R., Dorfman, A. H., and Royall, R. M., 2000. *Finite Population Sampling and Inference: A Prediction Approach*, John Wiley & Sons, New York.
- Vogelmann, J. E., Howard, S. M., Yang, L., Larson, C. R., Wylie, B. K., and Van Driel, N., 2001. Completion of the 1990s national land cover data set for the conterminous United States from Landsat Thematic mapper data and ancillary data sources, *Photogrammetric Engineering & Remote Sensing*, 67:650-662.
- Volstad, J. H., Roth, N. E., Mercurio, G., Southerland, M. T., and Strebel, D. E., 2003. Using environmental stressor information to predict the ecological status of Maryland non-tidal streams as measured by biological indicators, *Environmental Monitoring and Assessment*, 84:219-242.
- White, R., 2006. Pattern based map comparisons, *Journal of Geographical Systems*, 8:145-164.
- Wickham, J. D., Stehman, S. V., Smith, J. H., and Yang, L., 2004a, Thematic accuracy of the 1992 National Land-cover Data for the western United States, *Remote Sensing of Environment*, 91:452-468.
- Wickham, J. D., Stehman, S. V., Smith, J. H., Wade, T. G., and Yang, L., 2004b, A priori evaluation of

- two-stage cluster sampling for accuracy assessment of large-area land-cover maps, *International Journal of Remote Sensing*, 25:235-1252.
- Wulder, M. A., Franklin, S. E., White, J. C., Linke, J., and Magnussen, S., 2006. An accuracy assessment framework for large-area land cover classification products derived from medium-resolution satellite data, *International Journal of Remote Sensing*, 27:663-683.
- Zhu, Z., Yang, L., Stehman, S. V., and Czaplewski, R. L., 2000. Accuracy assessment for the U. S. Geological Survey regional land-cover mapping program: New York and New Jersey region, *Photogrammetric Engineering & Remote Sensing*, 66:1425-1435.

Table 1. NLCD 2001 Land-Cover Classes.

Level II	Level I
(11) Open Water	(10) Water
(12) Perennial Ice/Snow	
(21) Developed, Open Space	(20) Developed (Urban)
(22) Developed, Low Intensity	
(23) Developed, Medium Intensity	
(24) Developed, High Intensity	
(31) Barren Land	(30) Barren Land
(41) Deciduous Forest	(40) Forest
(42) Evergreen Forest	
(43) Mixed Forest	
(52) Shrub/Scrub	(50) Shrub/Scrub
(71) Grassland/Herbaceous	(70) Grassland/Herbaceous
(81) Hay/Pasture	(80) Agriculture
(82) Cultivated Crops	
(90) Woody Wetlands	(90) Wetland
(95) Emergent Herbaceous Wetlands	

Table 2. Applications of NLCD 1992

Source	Date	Description
Durbrow et al.	2001	develop conservation plan for southeast
Sheeder et al.	2002	estimate atmospheric nitrogen deposition
Heilman et al.	2002	national assessment of forest fragmentation
Heinz Center	2002	national environmental status report
Nolan et al.	2002	estimate groundwater nitrate contamination
Smith et al.	2002	track environmental impact of small water bodies
Riitters et al.	2002	national assessment of forest fragmentation
Amundson et al.	2003	national assessment of soil disturbance
Mayer and Cameron	2003	estimate changes in breeding bird populations
DeLiberty and Legates	2003	long-term soil moisture trends in Oklahoma
Volstad et al.	2003	identify impaired streams in Maryland
Lawler et al.	2003	select conservation area site
Alexander et al.	2004	estimate total phosphorus loads for Mississippi River watershed
Treves et al.	2004	predict likelihood of wolf predation on livestock
SC DHEC ¹	2004	determine fecal coliform TMDL for Big Wateree Creek, SC
Marshall et al.	2004	show changes in sensible and latent heat flux.
Radeloff et al.	2005	national estimate of wildland-urban interface

¹ Development of TMDLs (Total Maximum Daily Loads) is a common application of NLCD 1992.

Under the Clean Water Act (P.L. 92-500), TMDLs are required when water bodies are identified as not meeting water-quality standards. A search using the keywords NLCD and TMDL in *Google* identified 770 “hits.” Routine use of NLCD for TMDLs includes watershed characterization of land-cover composition and development of loading functions. The South Carolina Department of Health and Environmental Control (SC DHEC) TMDL is listed to represent this common use of the data.

Table 3. NLCD 2001 Accuracy Assessment Objectives

1. *Assess accuracy of NLCD 2001 for Level I and II land-cover classes via a per-pixel, error matrix approach, reporting overall, user's and producer's accuracies.* The MRLC partners ranked this objective as the highest priority because it describes key accuracy information in an intuitive and informative format. This objective is also important to determine if the mapping methods employed to create NLCD 2001 improved product quality relative to NLCD 1992.
2. *Assess accuracy of percent urban imperviousness and percent tree canopy.* This is a high priority objective because of the anticipated wide use of these NLCD 2001 databases. For example, percent impervious surface is a key environmental variable for studies of water quality and quantity (Brabec *et al.* 2002), and percent tree canopy is a key environmental variable in fire modeling (Schmidt *et al.* 2002).
3. *Assess accuracy of Level I land-cover composition at one or more spatial supports.* A very common use of the NLCD 1992 data was estimation of land-cover proportions of the Level I classes for larger spatial supports (e.g., watersheds) to drive models and assessments (Table 2). Similar use of NLCD 2001 is anticipated, thus making accuracy assessment of Level I land-cover composition a high priority.
4. *Assess accuracy of Level I net change derived from NLCD 2001 and NLCD 1992.* Because changes in data and mapping methods between NLCD 1992 and 2001 do not currently support change detection, MRLC is in the process of producing "retrofit" NLCD 1992 to support change detection. As the retrofit data become available, change accuracy assessment will focus on net change - aggregate change by class over a specified unit of support (Stehman and Wickham 2006). Net change is consistent with the way in which many users of NLCD will apply the change data. Assessing the accuracy of a per-pixel, gross change product (e.g. Biging *et al.* 1998) is not warranted until a time series of data using unchanging (or nearly so) mapping methods is available.

5. *Conduct analyses that take into account reference data error.* Reference data are subject to errors of various origins (Congalton and Green 1993, Powell *et al.* 2004), and accuracy results should be produced that reflect the potential impact of these errors. Hagen (2003) provides a rigorous analysis framework employing concepts of “fuzziness of location” and “fuzziness of category” that is applicable to this objective.

6. *Assess accuracy of landscape features and patterns.* Expanding beyond a per-pixel assessment to evaluate “higher order properties” of maps such as area and shape of land-cover patches and landscape pattern metrics would be desirable. The nascent state of development of practical methods to achieve this objective contributes to its low priority.

Table 4. NLCD 2001 Desirable Sampling Design Criteria

1. *Probability sample*: A probability sampling design should be implemented to ensure objectivity and to provide a rigorous underlying statistical basis via the well-established framework of design-based inference (Särndal *et al.* 1992).
2. *Adequately precise*: The design should result in adequate precision (i.e. small standard errors) for the accuracy estimates addressing priority objectives. Precision is determined by sample size and the sampling strategy implemented, along with the inherent variability of the population being sampled.
3. *Cost effective*: Cost should be minimized without sacrificing statistical rigor or quality. Cost is primarily a function of sample size, and sample size depends on the objectives the assessment is designed to address along with the precision specified for the estimates (Stehman 2001). Reducing the number of objectives and relaxing the precision requirements for each estimate are options for lowering cost. Improving efficiency of the sampling design may also reduce cost.
4. *Nationally uniform*: The design should be nationally uniform to simplify analyses and communication of design methodology and results. Design features that should be uniformly maintained include the basic selection protocols and the choice of sampling units. For example, if a stratified design is implemented, common strata (e.g. Level II land-cover classes) should be consistent nationally. If cluster sampling is implemented, the cluster size should be the same for each region.
5. *Geographically well distributed*: In addition to its intuitive appeal, a geographically well distributed sample is advantageous because it enhances options for regional reporting of results, distributes the workload uniformly by region, and may improve precision of the national level accuracy estimates.

6. *Accommodates analyses quantifying the impact of location error*: The sample should provide information contributing to understanding the impact of location error (i.e. spatial misregistration between the map and reference data) on the accuracy estimates. This criterion links directly to Objective 5.

Table 5. Probability Sampling Designs Employed for Large-Area Accuracy Assessments

Map Assessed	Strata	Clusters	General Design Description
NLCD 1992 (Stehman <i>et al.</i> 2003; Wickham <i>et al.</i> 2004)	a) 10 EPA Federal regions (geographic), b) land cover	6 km by 6 km areal unit (approximate size of NAPP photo)	Pixels within first-stage sample of areal units stratified by map land-cover class, followed by stratified random sample of pixels (implemented independently for each EPA Federal region)
IGBP DISCover (Scepan 1999)	Land cover	None	Stratified random sample
Utah GAP (Edwards <i>et al.</i> 1998)	Combination of three ecoregions by two “access” attributes (road, off-road)	a) USGS quadrangle, b) linear cluster of 10 sample points (off-road stratum only)	100 quadrangles selected via probability sampling design, road and off-road strata delineated within each sample quadrangle, followed by random sample of 10 points in the road stratum and one 10 point linear cluster in the off-road stratum per quadrangle sampled
Iowa GAP pilot study (Nusser and Klaas 2003)	a) geographic, b) land cover	7.5’ USGS quadrangle sheet	Systematic sample of quadrangles within geographic strata, pixels within the sampled quadrangles stratified by land cover class, stratified random sample of these

			pixels (3x3 block around each sample pixel visited for obtaining ground condition)
Land Cover Map of Great Britain (Fuller <i>et al.</i> 1998)	Environmental regions	1 km by 1 km block	Stratified random sample of the 1 km ² blocks
Global Land Cover (GLC) 2000 (Mayaux <i>et al.</i> 2006)	Combination of high and low priority land cover classes, and high and low diversity of land cover (4 strata)	Landsat scenes (intermediate step of defining Voronoi polygons around the center point of Landsat scenes)	A stratified systematic sample of Landsat scenes was selected, and five 3 km by 3 km blocks were chosen according to a fixed pattern within each sampled scene
Land cover map of forested area in Canada (Wulder <i>et al.</i> 2006)	a) 10 forested terrestrial ecozones of Canada, b) land cover	1:250000 National Topographic System (NTS) mapsheet (approximately 145 km by 111 km)	First-stage sample of NTS mapsheets, followed by stratifying by land cover class all pixels within the sample mapsheets, and a simple random sample of 100 pixels per class per mapsheet

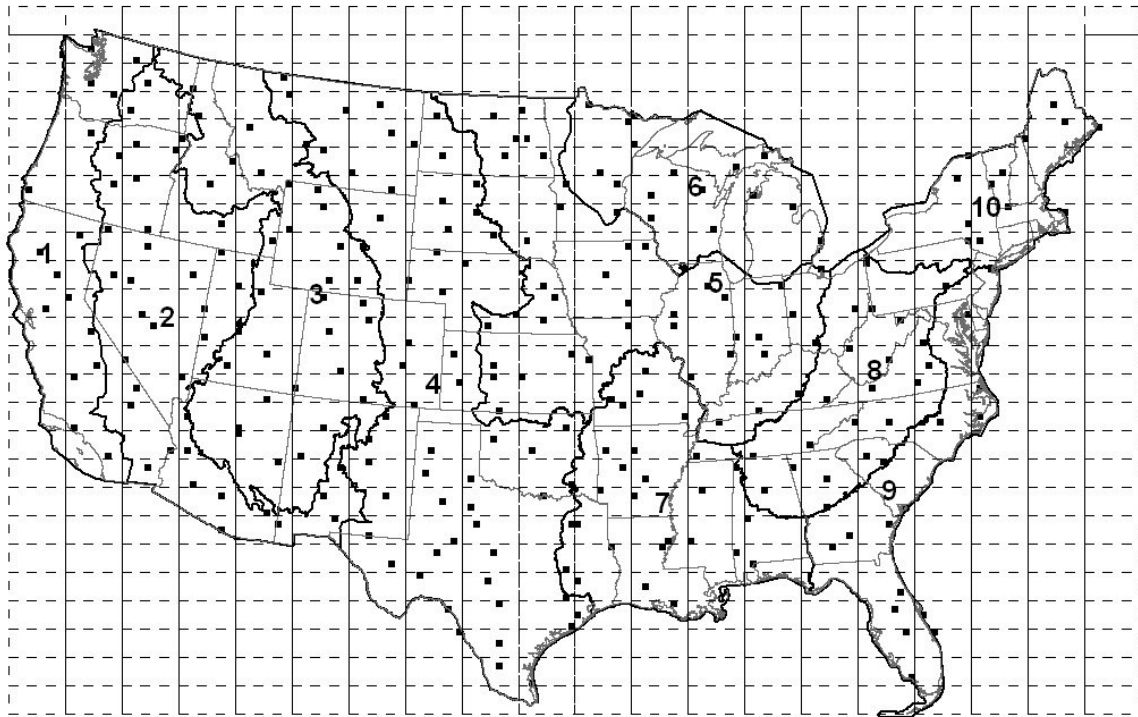


Figure 1: First-stage sample for NLCD 2001 accuracy assessment. The solid lines delineate the 10 accuracy assessment regions, and the dashed lines delineate the geographic strata (240 km by 120 km). The first-stage sample selects one 24 km by 24 km primary sampling unit within each geographic stratum (solid block). The second-stage sample is stratified by the Level II land-cover classes within each first-stage sample PSU, and a stratified random sample of pixels selected for each accuracy assessment region. The initial phase of the second-stage design will be implemented in Region 8.