Assessing Accuracy of Land-Cover Change Data Aggregated to a Fixed Spatial Support

Stephen V. Stehman

State University of New York College of Environmental Science and Forestry
320 Bray Hall, Syracuse, NY 13210 USA
Telephone: +1 315 470 6692
Email: svstehma@syr.edu

and

James D. Wickham

United States Environmental Protection Agency, Landscape Characterization Branch
E243-05, Research Triangle Park, NC 27711 USA
Telephone: +1 919 541 3077
Email: Wickham.James@epamail.epa.gov

Abstract

A method for assessing the accuracy of land-cover change products is described that evaluates net change at a specified spatial support. Accuracy is quantified by the mean absolute deviation, a metric derived from the absolute value of the difference between the map change in a land-cover class and the true change in the land-cover class. Several sampling design options are described for collecting the reference data. Stratification is a key design feature because of the desire to increase sample size in the rare high change areas of the map. A protocol for *a priori* evaluation of different sampling designs is described. Details of the accuracy assessment protocol are illustrated via a change product derived from the 1992 and preliminary 2001 National Land-Cover Data (NLCD).

1. Introduction

Land-cover change data are valuable to describe patterns in the environment, and they are often incorporated as driving or explanatory variables in environmental modeling. Applications of land-cover change data may require spatially explicit information in the form of a map, or use the data aggregated to some spatial unit to investigate relationships at a designated spatial support (here we use Dungan et al.'s (2003) definition of *support* as a property of a variable related to analyzing or modeling data). Although the spatial support can be as small as a point, our focus here is on larger areas ranging from 1 ha to several hundred square kilometers because environmental assessment and modeling studies often aggregate land-cover measured for individual pixels. The spatial unit employed in the assessment may be a regular geometric shape such as a square or circle, or an irregular polygon such as a county, province, or watershed. For example, changes in NDVI over a twenty-year period were compiled by watershed for an environmental assessment of the mid-Atlantic states (Jones et al., 1997), because watersheds are a logical and intuitive unit on which to base environmental management decisions. Likewise, tessellation of a regular spatial unit (e.g., 3x3, 5x5 km cells) has been used to examine geographic patterns of the impact of land-cover change on nutrient dynamics and breeding bird habitats (Jones et al., 2001; Wickham et al., 2002). The watershed- and tessellation-based assessments are similar in their philosophy in that it is often necessary to summarize change at the level of individual pixels to render them meaningful for environmental assessments and modeling studies.

The objective of this article is to describe a general sampling design and analysis strategy for assessing the accuracy of land-cover change at spatial scales larger than a single pixel or minimum mapping unit (mmu) that treats small groups of adjacent pixels as homogeneous units. Use of larger assessment units (e.g., counties, watersheds) represents a fundamental departure from land-cover change accuracy assessments based on single pixels or a mmu. Maintenance of homogeneity is not possible for larger assessment units, and hence crisp “from” and “to” land-cover labels cannot be applied. A county or a watershed does not change “from” forest “to” urban homogeneously. Four major themes are addressed: metrics to quantify accuracy of change, formulas for estimating these accuracy metrics, sampling design, and methods for evaluating the anticipated design performance when planning the assessment (i.e., *a priori* design evaluation). The assessment strategy is illustrated via an example in which the accuracy of change data derived from the 1992 National Land-Cover Data (NLCD) of the United States (Vogelmann et al., 2001) and a preliminary version of the 2001 NLCD. Users may be tempted to simply overlay the two NLCD products pixel by pixel to derive change. This use of NLCD data is discouraged because it likely will result in large amounts of erroneous mapped change due to inexact spatial registration of the two products and other problems associated with post-classification change detection. Although a per-pixel change map is not an intended use of NLCD data, it is anticipated that by spatially aggregating the NLCD data, the change values will prove useful for applications employing the data at larger spatial supports.

The methodology we propose is different in objectives and implementation from the traditional approach to accuracy assessment of land-cover change described in Biging et al. (1998) and Congalton and Green (1999). An example helps to highlight the differences. Suppose the original change map consists of 30-m pixels. The traditional approach to accuracy assessment is based on an error matrix and measures of accuracy derived from this matrix. For example, the proportion of those pixels mapped as changing from forest to urban that in reality have changed from forest to urban is the user's accuracy for forest to urban change. The change accuracy error matrix is large even for a small number of land-cover classes because the error matrix shows all possible transitions (i.e., “from-to” changes) between classes, as well as a “no change” category for each land-cover class.

The assessment of net change is designed to address different objectives from the traditional approach. First, the spatial support to which the assessment is intended is specified (e.g., a 20 by 20 pixel block). The support is dictated by the application(s) for which the map will most likely be used. Accuracy is then defined in terms of the map change and the true change measured on spatial units corresponding to this spatial support. For example, to assess accuracy of forest change, the mapped area of forest for both time t1 and time t2 for each spatial unit is measured, and net change is the difference between these values. Map net change is then compared to the true net change of each spatial unit. Similar comparisons are conducted for each land-cover class. This approach evaluates change aggregated to the predetermined spatial support, and compensating errors within the assessment unit may offset each other (e.g., a misclassification of one pixel as forest-to-urban change may be offset by a misclassification of one pixel as urban-to-forest change). In contrast, the traditional approach focuses on location-specific accuracy, addressing the question of whether each individual pixel's change status is mapped correctly. Assessing accuracy for aggregations of pixels is sometimes labeled "non-site-specific" accuracy, a label that often connotes aggregation over the entire region mapped. The assessment we propose is "non-site-specific" to a degree, but also captures some spatially explicit character of accuracy at the scale of the spatial unit defined by the support.

Another fundamental difference in the approaches is the definition of change assessed, net change in our approach versus gross change in the traditional approach. Net change is change at an aggregate level (Fuller, 1999, p. 336), for example the difference in the area of forest between time t1 and time t2. Net change does not focus on the internal, location-specific mechanisms of how change takes place. For example, a net loss of 10% forest could be the result of losing 10% of the forest area to urban land cover, losing 5% of the forest area to agriculture and 5% to urban, or losing 25% of the forest area to agriculture offset by a 15% gain in forest attributable to change from agriculture to forest. The traditional approach to accuracy assessment focuses on gross change. Gross change is defined at the individual pixel level, and addresses the specific land-cover transitions of change. A pixel-by-pixel change map represents gross change. For example, gross change would quantify the number of pixels changing from forest to agriculture. The "from-to" change error matrix advocated by Congalton and Green (1999) in the traditional approach to change accuracy assessment reflects the focus on gross change.

2. Accuracy Assessment Protocol

Change accuracy assessment requires the three basic components common to accuracy assessment of a single point in time. These components are the response design for characterizing the ground or reference condition, the sampling design, and the analysis (estimation formulas). Each of these components will be described for the aggregated change accuracy assessment.

2.1 Response design

The response design includes the protocols for determining the true change on the ground, as well as methods for defining agreement between the true change and the map change. "Reference data" is used to describe the change data characterizing the ground condition. These reference data likely also contain errors, but they are assumed to represent a more accurate characterization of reality than the map itself. The response design requires choosing the spatial unit upon which the assessment will be based. The assessment unit may be a fixed-area plot such as a square, rectangle, or circle, or an irregular unit such as a watershed or county. These units should partition the region assessed, and they should retain their identity for both dates (i.e., the same set of units is applicable to both time t1 and time t2). The choice of spatial unit obviously strongly influences the assessment, and the decision is complicated by the likelihood that users may be interested in different spatial supports.

Our example analyses based on the NLCD change data employ a 20x20 pixel (36 ha) assessment unit. This unit was chosen because it is large enough to diminish some of the effect on estimates of net change attributable to misregistration when overlaying the two NLCD products, it is a manageable size for obtaining reference data, and it is of sufficient size to be relevant for applications employing the change data. For each assessment unit, the area or percent of area occupied by each land-cover class is obtained from the 1992 and 2001

NLCD. Net change for each land-cover class is then the difference between the 2001 and 1992 values.

2.2 Quantifying accuracy

The descriptive results of the accuracy assessment are organized by “reporting domains" of change. These domains are defined by the magnitude and direction of change for each land-cover type. A reporting domain is a subset of the full region, and it is defined to enhance the interpretive value of the description. A domain consists of all spatial units in the region that meet the defining conditions of the domain, and a spatial unit may belong to more than one reporting domain. For example, one set of reporting domains for forest change could be forest loss of 15% or more (>–15% change), 7.5 to 15% forest loss (-7.5% to –15% change), 2.5% to 7.5% loss, 2.5% loss to 2.5% gain (-2.5% to 2.5%), 2.5% to 7.5% gain in forest, 7.5% to 15% gain, and greater than 15% gain in forest. The reporting domains defined in this example set include very high change domains, and thus target a scenario in which map high change values are anticipated. Reporting domains can be defined or revised after the data are collected, and it is possible that the domains defined for one land-cover type differ from the domains of another land-cover type.

It is important to distinguish reporting domains from strata employed in the sampling design. Reporting domains do not impact how the sample is selected, whereas strata directly influence the sampling units chosen. Strata constitute a fixed structure of the sampling design, whereas reporting domains are a characteristic of the analysis. A sampling unit can belong to several reporting domains, but must belong to exactly one stratum.

The primary descriptor of change accuracy emphasized in this article is the mean absolute deviation, MAD, computed for each reporting domain. The notation required is as follows. Let m_u denote the net change value derived from the map for sampling unit (block) u, and r_u denote the net change value derived from the reference data, with the difference $d_u=r_u-m_u$. MAD is the average of $|d_u|$ for all units in the reporting domain within the region. For example, for the forest class, MAD is computed for each reporting domain to quantify average absolute disagreement between the map net change and reference net change of forest.

2.3 Sampling design

In practice, the region mapped will need to be sampled and estimates of the accuracy metrics derived from the sample data. The ultimate sampling unit is the spatial unit (support) defined for the evaluation (e.g., a 20x20 pixel block in the NLCD example). For those reporting domains that are of high interest but represent a relatively rare condition, stratification may be implemented to ensure MAD estimates are acceptably precise for these domains. In the NLCD change assessment, the rare domains of greatest interest are the high change domains (either gain or loss). Do these areas of high map change truly represent high change on the ground?

The focus on high change strata creates a complication in the stratum assignment process. Some elements (spatial units) of the population may meet conditions for membership in more than one stratum. For example, if both a high loss forest stratum and a high gain urban stratum exist, it is entirely possible that some elements of the population will have both high forest loss and high urban gain, and thus are candidates for two different strata. A requirement of stratified sampling is that each element must belong to exactly one stratum, so the assignment protocol must guarantee that this requirement is met. The protocol we invoke is sequential, assigning each element to a stratum via a pre-determined order for checking if an element meets the conditions defining each stratum. An example is provided in section 3.2.

Once the strata have been identified, the sample allocation to strata must be decided. When the objective is to obtain precise estimates for each stratum, conventional guidelines for stratified sampling suggest allocating larger sample sizes to the more variable strata and to the more important strata. The emphasis on reporting domains to characterize accuracy requires recognizing an additional dimension of the sample allocation decision. Because many of the reporting domains are not defined as strata, stratification may actually diminish precision of these domain estimates relative to simple random sampling (SRS). That is, while the stratified design will improve precision for those domains targeted as strata, it may be detrimental to precision of estimates for other domains. This feature is illustrated in the numerical examples provided later. Consequently, choosing strata and allocating sample size to strata must consider not only will these choices improve precision for some domains, but also how much harm stratification creates for precision of other reporting domains.

3. Example Assessment Using NLCD Change Data

We illustrate the accuracy assessment protocol via application to a change product consisting of 10,000, 20x20 pixel blocks, with each pixel 30 m per side. These 10,000 blocks represent a simple random sample from a larger population of 180,000 such blocks located in the mid-Atlantic region of the United States (Figure 1). The map labels assigned to each pixel in this population are derived from the 1992 NLCD and a preliminary version of the 2001 NLCD. For some of the illustrative analyses, we also employ hypothetical reference data derived from the change between the 2001 NLCD and 1989 National Oceanographic and Atmospheric Administration (NOAA) Coastal Change and Analysis Program (CCAP) data.

The NLCD (Vogelmann et al., 2001) and NOAA CCAP (Dobson et al., 1995) land-cover data sets were grouped into a simple classification scheme for the comparison of map and reference change (Table 1). The simple classification scheme approximates the Anderson et al. (1976) Level I detail. Grouping into a simpler classification scheme was done in part to reconcile the differences between CCAP and NLCD at the more detailed Level II classification, and because there is little compelling reason for change detection at Level II. Environmental managers and land use planners are less interested in change within different types of forest or urban, for example, than changes between forest, agriculture, urban, and wetland. Forested wetlands in all three data sets were grouped with upland forest in the generalized scheme because previous NLCD accuracy assessments indicated that reference data for wetland forests were nearly as likely to be labeled as upland as wetland (Yang et al., 2001; Stehman et al., 2003). At a generalized level, a pixel labeled as wetland forest would be considered correct if the reference label was either the wetland or forest class. The most significant disagreement between NOAA and NLCD generalized classes occurs with the agricultural label. The NOAA grassland class (detailed legend) includes agricultural pasturelands as well as golf courses and the grass infields surrounding airport runways. The NLCD classification scheme attempts to differentiate between pasturelands and other grass-covered land uses. As a result, some areas labeled agriculture in the NOAA data will be labeled urban in the NLCD data. These conceptual differences will be reflected quantitatively in the "map" versus "reference" comparisons.

3.1 Accuracy of net change for the NLCD product

Table 2 displays the mean and median absolute deviations by reporting domain for forest, agriculture, and urban net change derived from the NLCD (recall that the 2001 NLCD and 1989 CCAP land-cover maps are used as the hypothetical reference data in this example).

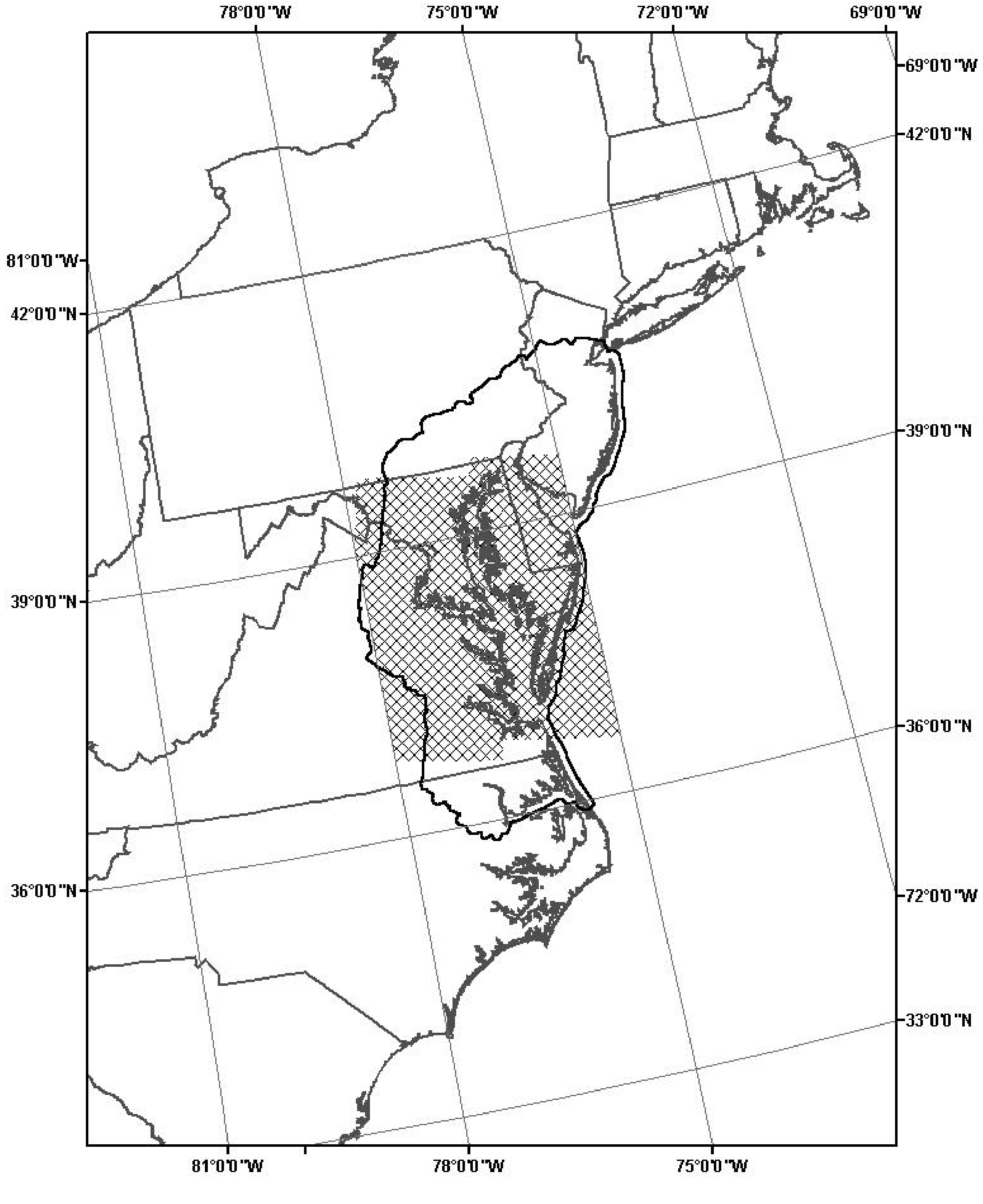

Figure 1. Mid-Atlantic region of the United States used in the example net change accuracy assessment. The cross-hatched region is the area from which the 10,000 20x20 pixel blocks were selected.

Table 1. Grouping of Land-cover Classes

NLCD 1992 Classes		NLCD 2001 Classes		NOAA 1989 Classes	
Detailed	General	Detailed	General	Detailed	General
Open water	Water	Open Water	Water	High intensity developed	Urban
Low-density residential	Urban	Developed Open Space	Urban	High intensity developed	Urban
High-density residential	Urban	Developed low intensity	Urban	Cultivated land	Agriculture
Commercial, Industrial, Transportation	Urban	Developed medium intensity	Urban	Grassland	Agriculture
Bare rock, sand, clay	Barren	Developed high intensity	Urban	Deciduous Forest	Forest
Mining	Urban	Natural Barren	Barren	Evergreen Forest	Forest
Transitional[1]	No data	Deciduous Forest	Forest	Mixed Forest	Forest
Deciduous Forest	Forest	Evergreen Forest	Forest	Scrub/shrub	Forest
Evergreen Forest	Forest	Mixed Forest	Forest	Palustrine forested wetland	Forest
Mixed Forest	Forest	Hay/Pasture	Agriculture	Palustrine scrub/shrub wetland	Forest
Hay/Pasture	Agriculture	Row crops	Agriculture	Palustrine emergent wetland	Wetland
Row crops	Agriculture	Woody wetland	Forest	Estuarine emergent wetland	Wetland
Urban, recreational grasses	Urban	Emergent Wetland	Wetland	Unconsolidated shore	Barren
Woody wetland	Forest			Bare land	Barren
Emergent Wetland	Wetland			Water	Water

[1] Pixels labeled transitional in the NLCD 1992 data were converted to “no data” in all data sets.

Table 2. Mean and Median Absolute Deviations by Domains of Forest, Agriculture, and Urban Change for a Population of N=10,000 20x20 pixel blocks. The domains are defined by the NLCD net change for each class, and M_k denotes the number of 20x20 pixel blocks in the domain.

		Forest			Agriculture			Urban	
Domain	MAD	Median	M_k	MAD	Median	M_k	MAD	Median	M_k
1	0.116	0.085	1606	0.144	0.123	380	0.241	0.203	139
2	0.075	0.063	1734	0.099	0.085	616	0.131	0.105	238
3	0.060	0.040	1999	0.076	0.050	1112	0.081	0.058	527
4	0.044	0.020	2802	0.047	0.020	3323	0.015	0.003	7565
5	0.075	0.053	979	0.066	0.040	2171	0.045	0.020	827
6	0.104	0.083	560	0.084	0.065	1481	0.071	0.048	377
7	0.140	0.126	320	0.131	0.093	917	0.097	0.063	327

Domains (net change):
1: >-15% (more than 15% loss)
2: -15% to -7.5%
3: -7.5%to -2.5%
4: -2.5% to 2.5%
5: 2.5% to 7.5%
6: 7.5% to 15%
7: >15% (more than 15% gain)

The results generally show a pattern of increasing disagreement moving from the low change to the high change domains. The mean absolute deviation is generally higher than the median absolute deviation, indicating a right skewed distribution of absolute deviations, as well as potential extreme high values of $|d_u|$. The large values for mean and median absolute deviation suggest fairly substantial disagreements between the map and reference change condition. For example, MAD is 0.116 for forest domain 1 (forest loss exceeding 15%), indicating that the reference change differs (in absolute value) from the map change by nearly 12% on average. This large disagreement is not surprising given the preliminary nature of the 2001 NLCD product, and the fact that errors in the NOAA CCAP 1989 land-cover data will also contribute to disagreement. Table 3 displays mean and median raw deviations (i.e., not absolute value). The pattern of these differences is also intuitively reasonable, with large negative values occurring for domain 1 (high loss), followed by a monotonic increase in the deviations, ultimately peaking at the large deviations of domain 7 (high gain). That is, where the map displays large losses of a land-cover class, the reference data (on average) show that the loss was not as severe as the map indicates, thus the negative mean or median difference for the low domains. The converse occurs when map change shows high gains in a class.

Table 3. Mean and Median Deviations by Domains of Forest, Agriculture, and Urban Change for a Population of N=10,000 20x20 pixel blocks. Domains 1-7 are defined (see Table 2) by the NLCD net change for each class, and M_k denotes the number of 20x20 pixel blocks in the domain.

		Forest			Agriculture			Urban	
Domain	Mean	Median	M_k	Mean	Median	M_k	Mean	Median	M_k
1	-0.088	-0.065	1606	-0.110	-0.113	380	-0.232	-0.203	139
2	-0.046	-0.048	1734	-0.052	-0.068	616	-0.115	-0.099	238
3	-0.029	-0.023	1999	-0.007	-0.028	1112	-0.048	-0.040	527
4	-0.005	0.000	2802	0.019	0.000	3323	0.003	0.000	7565
5	0.027	0.038	979	0.046	0.025	2171	0.007	0.005	827
6	0.056	0.070	560	0.066	0.053	1481	0.036	0.030	377
7	0.105	0.111	320	0.116	0.083	917	0.060	0.038	327

3.2 Sampling design for NLCD change example

The sampling design evaluated is a stratified random sample, with the strata defined based on the map net change. A sequential, one-stratum-at-a-time approach was adopted to assign blocks to strata to satisfy the requirement that each areal unit (20x20 pixel block) in the region must be assigned to exactly one stratum. At each step, each block is checked to determine if it should be assigned to the stratum of record at that step. If the block meets the conditions defining the stratum, the block is placed in that stratum and removed from consideration as a member of any subsequent stratum. Those blocks not assigned at this step of the sequence nor assigned at a previous step are passed on to the next step. Suppose seven strata are defined, with the stratum assignment sequence as follows: 1) 15% forest loss, 2) 15% forest gain, 3) 15% urban loss, 4) 15% urban gain, 5) 15% agriculture loss, 6) 15% agriculture gain, and 7) all blocks not assigned in the first six steps (catch-all stratum consisting of low and moderate change blocks). The stratum assignment protocol is applied to all blocks in the mapped region (population).

The sequential approach represents one solution to the problem of how to assign blocks to strata when some blocks may satisfy conditions for membership in more than one stratum. For example, it is possible for a block to have high forest loss and high urban gain (e.g., forest clearing for residential development). Depending on the order specified for stratum assignment, a block could be designated to the high loss forest stratum (stratum 1) or the high gain urban stratum (stratum 4). In the example sequence, such a block would be assigned to stratum 1, the high forest loss stratum. As long as each block belongs to just one stratum, we are free to choose the assignment sequence.

Numerous other stratification options could be created. One option is to define strata based on a cross-classification of change types, as for example, by defining a stratum as having 15% loss in Forest and 15% gain in Urban. Alternatives to the sequential assignment protocol could be envisioned to accommodate blocks that meet conditions of more than one stratum. For example, we could randomly assign the block to one stratum or the other so that half the time such a block is assigned to each stratum. The current sequential procedure assigns all such blocks to the same stratum.

3.3 Estimation for the stratified design

The formulas and theory for estimating MAD for reporting domains are found in Cochran (1977, Sec. 5A.14) and Sarndal et al. (1992, p. 394). Both texts recommend using a combined ratio estimator to estimate a domain mean when the domain cuts across stratum boundaries, as is typically the case for the approach described. Recall that d_u is the difference between the map change and reference change. The parameter (population value) for the mean absolute deviation for domain k is then $\sum_U | d_u |/M_k$, where U denotes the set of all M_k 20x20 pixel blocks in the region. For block u of stratum h, let $y_{hu}=|d_u|$ if block u is in domain k, $y_{hu}=0$, otherwise, and let $x_{hu}=1$ if block u is in domain k, and $x_{hu}=0$, otherwise. Note that both y_{hu} and x_{hu} are 0 for any block not in domain k. MAD for domain k can be expressed as $R_k=\sum_{h=1}^{H}\sum_U y_{hu} / \sum_{h=1}^{H}\sum_U x_{hu}$, where H is the number of strata. The numerator of R_k is the population total of $|d_u|$, and the denominator is simply M_k. Standard formulas for stratified sampling can readily be applied to estimate the numerator and denominator of this ratio,

$$\hat{R}_k = \frac{\sum_{h=1}^{H} N_h \bar{y}_h}{\sum_{h=1}^{H} N_h \bar{x}_h}. \tag{1}$$

The estimated variance of $\hat{R}_k$ is

$$\text{var}(\hat{R}_k) = (1/\hat{M}_k^2)\sum_{h=1}^{H} N_h^2 (1-n_h/N_h)(s_{y,h}^2 + \hat{R}_k^2 s_{x,h}^2 - 2\hat{R}_k s_{xy,h})/n_h), \tag{2}$$

where $s_{y,h}^2$ is the sample variance of y_{hu} in stratum h, $s_{x,h}^2$ is the sample variance of x_{hu} in stratum h, $s_{xy,h}$ is the sample covariance between y_{hu} and x_{hu} in stratum h, and $\hat{M}_k$ is the estimated number of blocks in the domain (i.e., the denominator of $\hat{R}_k$). All the statistics and variables in equations (1) and (2) are specific to domain k, but the subscript k is not used in most cases to simplify notation. For some domains, all values of y_{hu} and x_{hu} may be 0 for one or more strata because no sample blocks of domain k are present for that stratum. This does not create a problem because the contribution to $\hat{R}_k$ or to var($\hat{R}_k$) is zero for such strata. Lastly, if we replace $|d_u|$ with d_u, the formulas estimate the mean difference and the variance of the estimated mean difference (as opposed to mean absolute difference).

3.4 *A priori* evaluation of sampling design

In planning the sampling design, it is invaluable to evaluate the potential performance of one or more candidate designs. The primary information available at the planning stage is the map change derived from the 1992 and 2001 NLCD products (i.e., the map change product itself). This information can be used to define candidate change strata, and to evaluate design performance. The *a priori* evaluation of various design options focuses on precision (standard errors) of the estimates specified by the accuracy assessment objectives. In the NLCD example, the estimation objectives focus on the reporting domains for each of three land-cover classes, forest, urban, and agriculture.

Two analyses are proposed for the *a priori* evaluation of design performance. One is the expected sample size for each reporting domain of each land-cover type (i.e., the number of sample blocks expected to occur in the >15% forest loss domain, the number in the domain for less than 2.5% gain or loss, etc.). The expected sample size for a domain is computed by multiplying the proportion of the stratum occupied by that domain times the sample size for that stratum, and then summing the results over all strata. For example, suppose we have three domains and two strata, and a stratified random sample of n_h=100 blocks from each stratum will be selected. The stratum proportions of domains A, B, and C are 0.6, 0.3, and 0.1 in stratum 1, and 0.45, 0.5, and 0.05 in stratum 2. The anticipated sample sizes in each domain resulting from this design (contributions from strata 1 and 2 in parentheses) would be 105 (60 and 45) in domain A, 80 (30 and 50) in domain B, and 15 (10 and 5) in domain C. Because of the randomization built into the sampling protocol, these are expected (average) sample sizes over all possible samples. The outcome of an individual sample will vary from these average numbers.

The second analysis is much more detailed requiring construction of a hypothetical population of reference (i.e., true change) data. Once this population has been constructed, d_u is known for all blocks in the population, and the precision of the estimated MAD for each domain can be calculated via

$$V(\hat{R}_k) = (1/M_k^2)\sum_{h=1}^{H} N_h^2(1-n_h/N_h)(S_{y,h}^2 + R_k^2 S_{x,h}^2 - 2R_k^2 S_{xy,h})/n_h), \tag{3}$$

where $S_{y,h}^2$ is the population variance of y_{hu} in stratum h, $S_{x,h}^2$ is the population variance of x_{hu} in stratum h, $S_{xy,h}$ is the population covariance between y_{hu} and x_{hu} in stratum h, and R_k is the true MAD for domain k. V($\hat{R}_k$) represents the variability of the estimator $\hat{R}_k$ over all possible samples that could be selected from the population using the design under consideration. In practice, V($\hat{R}_k$) is estimated by var($\hat{R}_k$) using the sample of blocks actually selected for the assessment. At the planning stage, V($\hat{R}_k$) is the relevant quantity because the sample has yet to be chosen.

Because calculating V($\hat{R}_k$) requires knowing d_u for the population (census), the *a priori* design evaluation depends on constructing one or more hypothetical populations that represent a good approximation to the true change condition. That is, the goal is to create a population that reflects the true magnitude and spatial pattern of the map errors to the degree that the general findings of the *a priori* evaluation will reveal preferred design options. The multi-objective character of accuracy assessment requires us to examine precision for the suite of domain estimates. No design option is likely universally best for all domain estimates, so choosing a design requires considering which domain estimates are most important.

To illustrate the approach, we use the population of 10,000 blocks described earlier for which we have hypothetical reference data. Five stratification options are evaluated. These include different numbers of strata (H=5, 6, or 7), different assignment sequences of blocks to strata, and different choices of the strata themselves. The five options are listed below. The % area of gain or loss defining the strata is specified in parentheses for the first stratum and this % applies to all strata for that option. The “no change” stratum is always defined as net change between 2.5% loss and 2.5% gain (-2.5% to 2.5%).

Option A: 7 strata, with the assignment sequence urban gain (>15%), urban loss, forest gain, forest loss, agriculture gain, agriculture loss, and no change

Option B: 6 strata, with the assignment sequence forest gain (>15%), forest loss, urban gain, urban loss, agriculture loss, and no change

Option C: 6 strata, with the assignment sequence urban gain (>15%), urban loss, forest gain, forest loss, agriculture loss, and no change

Option D: 5 strata, with the assignment sequence urban gain (>15%), urban loss, forest gain, agriculture loss, and no change

Option E: 5 strata, with the assignment sequence as in option D, but gain and loss are defined as >10% instead of >15%.

Table 4 displays the anticipated sample sizes for the seven reporting domains for two of the stratified options. As a crude guideline, assume that a sample size of 40-50 blocks would produce adequate precision for a domain estimate. Because the forest domains are all relatively common in the region, stratification does not produce a marked difference in the distribution of the sample among the forest domains (Table 4). The main advantage of stratification for the forest domain estimates accrues to domains 6 and 7, where the sample size is increased from 22 and 13 to over 40 in both domains. The estimates for the urban domains benefit most from the stratified options. The no change urban domain (domain 4) dominates in the region, so SRS will result in 75% of the sample being no urban change blocks. Because both options A and D have urban domains 1 and 7 as the first two strata in the sequential selection, the design ensures a sample size of 50 in both domains. The stratified options do not increase the sample size above 30 in domains 2, 3, 5, and 6, so the stratified options primarily benefit only the rarest urban domains. The agriculture domains are affected similarly, with the rarest domain (domain 1) benefiting from its identification as one of the strata, but other domains not gaining much in sample size relative to SRS. In general, note that those domains not identified as strata have smaller expected sample sizes than would result from SRS. This is the crux of the characteristic that the stratified designs may have poorer precision for these domains than SRS.

Table 4. Expected sample sizes by domain for forest, agriculture, and urban for stratified sampling options A (StrA) and D (StrD), and simple random sampling (SRS). Total sample size is 400 for all three designs. Stratified option A has a sample size of n$_h$=50 per stratum except for the low change stratum which has 100, and stratified option D also has a sample size of n$_h$=50 per stratum with 200 samples in the low change stratum. Option A has 7 strata, and option D has 5 strata.

	Forest			Agriculture			Urban		
Domain	StrA	StrD	SRS	StrA	StrD	SRS	StrA	StrD	SRS
1	72	55	64	92	92	15	50	50	6
2	68	53	69	20	24	25	12	7	10
3	49	58	80	23	32	44	19	13	21
4	75	92	112	65	91	133	214	228	303
5	35	41	39	52	67	87	26	29	33
6	44	46	22	47	51	59	22	23	15
7	57	57	13	94	40	37	50	50	13

The precision resulting from each stratification option for the seven reporting domains of each land-cover class is shown in Table 5. Precision for SRS is used as the baseline for comparison, and the sample size is maintained at n=400 for all designs, both stratified and SRS. Negative values in Table 5 indicate a reporting domain for which the stratified design improves precision relative to SRS, and positive values represent domains in which SRS produces better estimates. The results of Table 5 should be examined for two key features: how much does stratification improve precision for the important high change domains (i.e., large negative values for domains 1 and 7), and how much worse than SRS are the stratified estimates for those domains not defined as strata (i.e., large positive values for domains 2 through 6)?

Stratification improves precision for those reporting domains identified as strata, but typically results in precision poorer than SRS for the remaining domain estimates. This is not an unexpected result. Because the strata are constructed to focus sampling effort on the high change areas, these strata are not necessarily a good choice for all domain estimates. Most domain estimates combine sample data from several strata, but these strata may not be internally homogeneous for all domain variables. Internal homogeneity is the condition conducive to precise estimation for stratified sampling. The design implications of these results are that strata should be chosen to focus effort on only the most important domain estimates, and the stratification chosen should not result in variances much higher than would be achieved by SRS for those domains not identified as strata.

The urban results provide a good illustration of the evaluation process. All five stratification options have two high urban change strata, one for high loss and one for high gain, representing domains 1 and 7. Because these two domains are identified as strata, the stratified options are better than SRS (note the negative numbers) except for one case, domain 7 for Option B. Option B's sequence differs from the other four in that the urban stratum assignments are made after blocks are assigned to the high change forest strata. Some of the high urban gain blocks (urban domain 7) are likely assigned to the high forest loss stratum, and this produces slightly poorer precision for urban domain 7 than the other four stratification options.

The results for forest domain 1 provide another interesting outcome. Relative to SRS, stratification does not improve precision for this domain because it is a large domain. For example, under option B, we ensure that exactly 50 sample blocks from forest domain 1 are selected. Because this domain contains approximately 16% of the blocks in the population, on average SRS of 400 blocks will produce 64 sample blocks from this domain. The larger expected SRS sample size (n=64) results in a more precise estimate than the stratified options (n=50). Therefore, a domain that includes a large proportion of the region's blocks (i.e., a common domain) will not require identification as a stratum because adequate sample size for such a domain will be produced by SRS.

Another important conclusion derived from this *a priori* evaluation is that the sample size in the No Change stratum (i.e., less than 2.5% gain or loss) should be large. This is a standard recommendation when many estimates are required, such as the multiple domain estimates employed to describe accuracy of change. Increasing the sample size in the large No Change stratum is tantamount to creating a large simple random sample. SRS is not tailored to enhance precision for any particular domain estimate, but in contrast to stratification, it does not diminish precision for those domains not identified as strata.

Table 5. Comparison of precision of domain estimators for various stratified sampling options. All designs have n=400 20x20 pixel sample blocks selected from the N=10,000 blocks comprising the population. The sample sizes are 50 per stratum, except for the "no change" stratum that has all remaining samples needed to make up the total of 400. The seven reporting domains are the same as in Table 2, and the 'All' row represents the estimates for the entire region. The column M_k shows the number of blocks in the domain. The SRS column is the standard error of the domain estimator under simple random sampling. All other columns display the difference between the standard error of SRS and the stratified option, with positive values indicating that SRS is more precise for that domain.

a) Forest Domains

			Stratified Sampling Option					
Dom	M_k	MAD	A	B	C	D	E	SRS
1	1606	0.116	0.001	0.002	0.001	0.005	0.005	0.014
2	1734	0.075	0.010	0.004	0.004	0.003	0.002	0.008
3	1999	0.060	0.011	0.005	0.005	0.003	0.003	0.009
4	2802	0.044	0.011	0.005	0.005	0.003	0.003	0.008
5	979	0.075	0.015	0.006	0.006	0.004	0.003	0.013
6	560	0.104	0.014	0.006	0.006	0.005	-0.002	0.021
7	320	0.140	-0.017	-0.016	-0.017	-0.017	-0.009	0.031
All	10000	0.073	0.002	0.001	0.001	0.001	0.001	0.005

b) Agriculture Domains

			Stratified Sampling Option					
Dom	M_k	MAD	A	B	C	D	E	SRS
1	380	0.144	-0.014	-0.015	-0.014	-0.015	-0.008	0.027
2	616	0.099	0.013	0.006	0.006	0.006	-0.001	0.017
3	1112	0.076	0.013	0.006	0.006	0.005	0.004	0.014
4	3323	0.047	0.008	0.004	0.003	0.003	0.002	0.007
5	2171	0.066	0.009	0.004	0.004	0.003	0.002	0.010
6	1481	0.084	0.007	0.004	0.003	0.004	0.003	0.011
7	917	0.131	-0.001	0.001	0.000	0.007	0.006	0.022
All	10000	0.075	0.002	0.001	0.001	0.001	0.001	0.005

c) Urban Domains

			Stratified Sampling Option					
Dom	Mk	MAD	A	B	C	D	E	SRS
1	139	0.241	-0.055	-0.049	-0.055	-0.055	-0.044	0.076
2	238	0.130	0.025	0.021	0.020	0.023	0.003	0.036
3	527	0.081	0.013	0.010	0.008	0.009	0.007	0.018
4	7565	0.015	0.003	0.003	0.001	0.001	0.001	0.002
5	827	0.045	0.011	0.012	0.005	0.004	0.003	0.012
6	377	0.071	0.012	0.018	0.007	0.008	0.001	0.019
7	327	0.097	-0.014	0.008	-0.014	-0.014	-0.009	0.027
All	10000	0.032	0.001	0.001	0.000	0.000	-0.000	0.003

4. Discussion

Several components of the net change accuracy assessment protocol merit additional development. The methods described in this article are based on a single spatial support, for example the 20x20 pixel block in the NLCD analyses. But applications of change products will undoubtedly span a range of spatial supports. Using the protocol outlined, it is possible to assess accuracy for spatial support smaller than that chosen to determine the original spatial unit of the assessment. That is, in our example NLCD assessment, it would be possible to evaluate a spatial support smaller than the 36 ha unit used as the sampling unit. Each 20x20 pixel block could be partitioned into 4 10x10 pixel blocks, 16 5x5 pixel blocks, or even 100 2x2 pixel blocks. The data analysis for the smaller support sizes requires one-stage cluster sampling formulas treating the smaller units (e.g., the 10x10 pixel blocks) within the larger original sampling units as secondary sampling units. The sampling design is not tailored to control precision of the estimates for these smaller spatial supports. Consequently, we should not expect the assessment to be as good as it would be had the design been created with this spatial support as the focus.

Another important question related to spatial support is how large of a spatial unit can be reasonably assessed. For example, if applications commonly employ a spatial support of 5 km by 5 km, is it possible to obtain quality reference data for units of this size? Interpreting change over such large areas may be problematic. One possibility we are exploring is the use of two-stage cluster sampling. Rather than interpret an entire sampling unit, we would employ a probability subsample of smaller secondary sampling units within each large unit to estimate net change of the larger unit. Multistage sampling may also accommodate the desire to assess multiple spatial supports.

The sampling design we have described is based on a fixed partition of the region into units corresponding to the spatial support chosen. This approach has two unsatisfying features. It requires eliminating small portions on the boundary of the study region to maintain complete 20x20 pixel blocks. This problem becomes more severe for larger support sizes. Secondly, the assessment is dependent on the partition selected (e.g., shifting the tessellation a small amount will change the results). Although it is likely that the differences resulting from such a shift would be small, ideally the assessment would be immune to the tessellation's origin. Employing a "floating" spatial unit would resolve these problems to some extent. In this approach, individual pixels would first be selected, and then a 20x20 block derived from this pixel becomes the sampling unit (e.g., use the sample pixel as the upper left corner of the 20x20 pixel block). The spatial units would thus not be fixed by a partition of the region, but rather would "float" within the region depending on the pixels selected. How to stratify the region when using a floating plot requires investigation. Two-phase sampling for stratification is a possible solution. A large first-phase sample would be selected and each spatial unit in this sample would be assigned to its appropriate stratum. A stratified subsample of the first-phase sample would then be chosen and the response design applied to this second-phase sample.

The sampling designs we have proposed for the aggregated accuracy assessment of change can be applied to simultaneously produce a traditional, site-specific assessment of gross change. The traditional assessment entails a much more data-intensive response design protocol to obtain the individual pixel gross change, but the sampling design protocols established for the net change assessment would be applicable to the traditional gross change accuracy analysis. To attempt such a dual-purpose assessment would exacerbate a major difficulty confronting accuracy assessments of any kind - accommodating the multiple objectives desired of the assessment. Adequately accomplishing many objectives requires implementing a more complex sampling design to achieve these goals with little added cost, or keeping the design simple, thereby requiring greater cost to achieve equivalent precision to the more complex design.

5. Summary

A sampling design and analysis strategy for assessing the accuracy of mapped net change was developed. This strategy recognizes that many applications of the change product will aggregate the data to some spatial support to investigate or model relationships between net change and various other phenomena. In common with one-point-in-time accuracy assessments, change accuracy assessments should be based on a probability sampling design to provide a rigorous statistical foundation. Several modifications from the traditional approach to accuracy assessment were developed. Rather than base the analysis on the massive change/no change error matrix, the mean absolute deviation (MAD) between map and true change serves as the basis for describing accuracy. MAD is computed for several user-defined reporting domains for each land-cover class of interest, and the flexibility to tailor reporting domains to a particular user's needs is an attractive feature. As a consequence of employing MAD as the primary descriptive accuracy metric, the options for defining strata differ from the traditional use of the map land cover classes as strata for a one-point-in-time assessment, or strata based on the change and no change map areas for a change accuracy assessment. The key information for stratification is still garnered from the change product itself, but the strata definitions are complicated by the fact that the spatial units (i.e., elements of the population) may meet conditions for membership in several strata. Lastly, we developed methods for evaluating the anticipated performance of the sampling design at the planning stage. This *a priori* evaluation provides quantitative information as well as qualitative insights on the relative merits of different designs, thus allowing a more informed choice of which strata and sample allocations are likely to be most effective. General guidelines derived from our *a priori* evaluation of an NLCD change product were to use a small number of strata and to allocate a large sample size to the large “no change” (-2.5% to 2.5%) stratum.

References

Anderson, J. F., Hardy, E. E., Roach, J. T., and Witmer, R. E., 1976, A land use and land cover classification system for use with remote sensor data, U.S. Geological Survey Professional Paper 964, U.S. Geological Survey, Washington, DC, 28 pp.

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.

Brown, D. G., Pijanowski, B. C., and Duh, J. D., 2000, Modeling the relationships between land use and land cover on private lands in the Upper Midwest, USA. *Journal of Environmental Management*, 59, 247-263.

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

Congalton, R. G., and Green, K., 1999, *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices* (CRC Press, Boca Raton, FL).

Dobson, J. E., Bright, E. A., Ferguson, R. L., Field, D. W., Wood, L. L., Haddad, K. D., Iredale, H., Jensen, J. R., Klemas, V. V., Orth, R. J., and Thomas, J. P., 1995, NOAA Coastal Change Analysis Program (C-CAP): guidance for regional implementation, NOAA Technical Report NMFS 123 (U.S. Department of Commerce, Seattle, Washington).

Dungan, J. L., Perry, J. N., Dale, M. R. T., Legendre, P., Citron-Pousy, S., Fortin, M.-J., Jakomulska, A., Miriti, M., and Rosenberg, M. S., 2002, A balanced view of scale in spatial statistical analysis, *Ecography*, 25, 626-640.

Fuller, W. A., 1999, Environmental surveys over time. *Journal of Agricultural, Biological, and Environmental Statistics*, 4, 331-345.

Jones, K. B., Neale, A. C., Wade, T. G., Wickham, J. D., Cross, C. L., Edmonds, C. M., Loveland, T. R., Nash, M. S., Riitters, K. H., and Smith, E. R., 2001, The consequences of landscape change on ecological resources: an assessment of the United States mid-Atlantic region, 1973-1993. *Ecosystem Health*, 7, 229-242.

Jones, K.B., Riitters, K.H., Wickham, J.D., Tankersley, R.D., O'Neill, R.V., Chaloud, D.C., Smith, E.R., and Neale, A.C., 1997, An Ecological Assessment of the U.S. Mid-Atlantic Region. EPA/600/R-97/130, Office of Research and Development, U.S. Environmental Protection Agency, Washington, DC.

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

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* (in press).

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 and Remote Sensing*, 67, 650-662.

Wickham, J. D., O'Neill, R. V., Riitters, K. H., Smith, E. R., Wade, T. G., and Jones, K. B., 2002, Geographic targeting of increases in nutrient export due to future urbanization. *Ecological Applications*, 12, 93-106.

Yang, L., Stehman, S. V., Smith, J. H., and Wickham, J. D., 2001, Short Communication: Thematic accuracy of MRLC land-cover for the eastern United States. *Remote Sensing of Environment*, 76, 418-422.