Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Thematic accuracy of the NLCD 2001 land cover for the conterminous United States

J.D. Wickham [a],*, S.V. Stehman [b], J.A. Fry [c], J.H. Smith [d], C.G. Homer [e]

[a] U.S. Environmental Protection Agency, Environmental Sciences Division, Research Triangle Park, NC 27711, USA
[b] SUNY College of Environmental Science and Forestry, 320 Bray Hall, 1 Forestry Drive, Syracuse, New York, 13210, USA
[c] Stinger Ghaffarian Technologies (SGT) Inc., U.S. Geological Survey, 47914 252nd St., Sioux Falls, SD 57198, USA
[d] U.S. Geological Survey, Geographic Analysis and Monitoring Program, 12201 Sunrise Valley Dr., 519 USGS National Center, Reston, VA, USA
[e] U.S. Geological Survey, EROS Data Center, 47914 252nd St., Sioux Falls SD 57198, USA

ARTICLE INFO

Article history:
Received 5 October 2009
Received in revised form 21 January 2010
Accepted 24 January 2010

Keywords:
Stratified sampling
Cluster sampling
Reference data error
NLCD 1992

ABSTRACT

The land-cover thematic accuracy of NLCD 2001 was assessed from a probability-sample of 15,000 pixels. Nationwide, NLCD 2001 overall Anderson Level II and Level I accuracies were 78.7% and 85.3%, respectively. By comparison, overall accuracies at Level II and Level I for the NLCD 1992 were 58% and 80%. Forest and cropland were two classes showing substantial improvements in accuracy in NLCD 2001 relative to NLCD 1992. NLCD 2001 forest and cropland user's accuracies were 87% and 82%, respectively, compared to 80% and 43% for NLCD 1992. Accuracy results are reported for 10 geographic regions of the United States, with regional overall accuracies ranging from 68% to 86% for Level II and from 79% to 91% at Level I. Geographic variation in class-specific accuracy was strongly associated with the phenomenon that regionally more abundant land-cover classes had higher accuracy. Accuracy estimates based on several definitions of agreement are reported to provide an indication of the potential impact of reference data error on accuracy. Drawing on our experience from two NLCD national accuracy assessments, we discuss the use of designs incorporating auxiliary data to more seamlessly quantify reference data quality as a means to further advance thematic map accuracy assessment.

Published by Elsevier Inc.

1. Introduction

The National Land Cover Database (NLCD), developed by the MultiResolution Land Characteristics (MRLC) Consortium (www.mrlc.gov) continues to be the primary source of land-cover data in the United States. The paper announcing MRLC's inaugural land-cover map, NLCD 1992 (Vogelmann et al., 2001), has been cited 320 times,[1] reflecting the widespread need for the data. NLCD 1992 has been used to study habitat loss (Hoekstra et al., 2005), conservation options (Carr et al., 2002; Weber, 2004; Weber et al., 2006), the contribution of land remote sensing to ecological study (Cohen & Goward, 2004), urban sprawl (Radeloff et al., 2005), forest fragmentation (Heilman et al., 2002; Riitters et al., 2002), nitrate contamination of groundwater (Nolan et al., 2002), water quality (Doherty & Johnston, 2003), land use impacts on precipitation patterns (Marshall et al., 2004) and net primary productivity (Milesi et al., 2003), human exposure to disease vectors (Jackson et al., 2006), model Total Maximum Daily Loads (TMDL) for the Clean Water Act (Wagner et al., 2007), and many other applications (Stehman et al., 2008).

MRLC recently completed a second NLCD database (NLCD 2001) (Homer et al., 2007). Although it is too early to assess the full impact of the data, Homer et al. (2007) have been cited 22 times[1] since its public release in late 2007. The widespread use of NLCD 1992 and the continuing need for nationwide land-cover data suggest that NLCD 2001 will be used as widely as its predecessor.

A nationwide land-cover accuracy assessment for NLCD 1992 was completed to support the use of those data (Stehman et al., 2003; Wickham et al., 2004). Here we document the methodology used to assess the accuracy of NLCD 2001 and report the conterminous national land-cover thematic accuracy results for NLCD 2001. Thematic accuracy assessment of the NLCD 2001 land-cover data was chosen as the top priority among many emerging accuracy assessment tasks that arise from the continued development of NLCD because of the widespread use of the land-cover data (Stehman et al., 2008).

2. Methods

The NLCD 2001 maps 16 land-cover classes (Table 1) across the conterminous United States at a nominal pixel scale of 30-m × 30-m with a minimum mapping unit of 5 pixels (see Homer et al., 2004, 2007 for full details of the classification procedures). Stehman et al. (2008) outlined the conceptual framework for the accuracy assessment of NLCD

* Corresponding author.
E-mail address: wickham.james@epa.gov (J.D. Wickham).

[1] The number of citations was based on a search at http://scholar.google.com, which reported 457 citations for Vogelmann et al. (2001) and 35 citations for Homer et al. (2001). Our tallies include only citations by other peer-reviewed articles. Searches were conducted on 08/21/2009.

0034-4257/$ – see front matter. Published by Elsevier Inc.
doi:10.1016/j.rse.2010.01.018

Table 1
NLCD 2001 land-cover classes (http://www.mrlc.gov/nlcd_definitions.php). Classes found in Alaska only are not included in this table, but are listed on the website. Class 12 was not included in the accuracy assessment (see text). Level I classes are represented by the tens digit of the numeric code (e.g., all classes with numeric codes in the 20s comprise the Level I urban class).

11. Open water—All areas of open water, generally with less than 25% cover of vegetation or soil.
12. Perennial ice/snow—All areas characterized by a perennial cover of ice and/or snow, generally greater than 25% of total cover.
21. Developed, open space—Includes areas with a mixture of some constructed materials, but mostly vegetation in the form of lawn grasses. Impervious surfaces account for less than 20% of total cover. These areas most commonly include large-lot single-family housing units, parks, golf courses, and vegetation planted in developed settings for recreation, erosion control, or aesthetic purposes.
22. Developed, low intensity—Includes areas with a mixture of constructed materials and vegetation. Impervious surfaces account for 20–49% of total cover. These areas most commonly include single-family housing units.
23. Developed, medium intensity—Includes areas with a mixture of constructed materials and vegetation. Impervious surfaces account for 50–79% of the total cover. These areas most commonly include single-family housing units.
24. Developed, high intensity—Includes highly developed areas where people reside or work in high numbers. Examples include apartment complexes, row houses, and commercial/industrial. Impervious surfaces account for 80 to 100% of the total cover.
31. Barren land (rock/sand/clay)—Barren areas of bedrock, desert pavement, scarps, talus, slides, volcanic material, glacial debris, sand dunes, strip mines, gravel pits, and other accumulations of earthen material. Generally, vegetation accounts for less than 15% of total cover.
41. Deciduous forest—Areas dominated by trees generally greater than 5 meters tall, and greater than 20% of total vegetation cover. More than 75% of the tree species shed foliage simultaneously in response to seasonal change.
42. Evergreen forest—Areas dominated by trees generally greater than 5 meters tall, and greater than 20% of total vegetation cover. More than 75% of the tree species maintain their leaves all year. Canopy is never without green foliage.
43. Mixed forest—Areas dominated by trees generally greater than 5 meters tall, and greater than 20% of total vegetation cover. Neither deciduous nor evergreen species are greater than 75% of total tree cover.
52. Shrub/scrub—Areas dominated by shrubs; less than 5 meters tall with shrub canopy typically greater than 20% of total vegetation. This class includes true shrubs, young trees in an early succession stage, or trees stunted from environmental conditions.
71. Grassland/herbaceous—Areas dominated by grammanoid or herbaceous vegetation, generally greater than 80% of total vegetation. These areas are not subject to intensive management such as tilling, but can be utilized for grazing.
81.Pasture/hay—Areas of grasses, legumes, or grass–legume mixtures planted for livestock grazing or the production of seed or hay crops, typically on a perennial cycle. Pasture/hay vegetation accounts for greater than 20% of total vegetation.
82. Cultivated crops—Areas used for the production of annual crops, such as corn, soybeans, vegetables, tobacco, and cotton, and also perennial woody crops such as orchards and vineyards. Crop vegetation accounts for greater than 20% of total vegetation. This class also includes all land being actively tilled.
90. Woody wetlands—Areas where forest or shrubland vegetation accounts for greater than 20% of vegetative cover and the soil or substrate is periodically saturated with or covered with water.
95. Emergent herbaceous wetlands—Areas where perennial herbaceous vegetation accounts for greater than 80% of vegetative cover and the soil or substrate is periodically saturated with or covered with water.

2001. The three major components of the accuracy assessment methodology, the sampling design, response design, and analysis (Stehman & Czaplewski, 1998) are described in this section.

2.1. Sampling design

The sampling design for obtaining the reference data was based on the design implemented for the NLCD 1992 accuracy assessment (Stehman et al., 2003). The sampling design was a two-stage cluster sample with three levels of stratification (Fig. 1). The first level of stratification was created by partitioning the conterminous United States into 10 geographic regions, which were constructed by aggregating the mapping zones used for NLCD 2001 (Homer et al., 2004). These 10 geographic strata facilitated regional reporting of accuracy and provided an indication of how accuracy varied spatially across the United States. The geographic stratification also ensured that the sample size allocated to regionally rare land-cover classes would be large enough to produce precise estimates of user's accuracy for these rare classes (Stehman et al., 2003). The land-cover composition of each of the 10 regional strata is provided in the Supplementary Material (Table S1). The sampling design for assessing the accuracy of NLCD 1992 also had 10 regional strata, but the strata for that assessment were administrative regions defined by the U.S. Environmental Protection Agency. We replaced the administrative regions used for the NLCD 1992 assessment with geographic strata for the NLCD 2001 assessment because of the correlation between class abundance and class accuracy.

Each of the 10 regions was then further partitioned into frame cells that were 120-km x120-km. The frame cells formed the second layer of stratification and may be viewed as geographic substrata within the 10 regional strata. A first-stage sample of 12-km × 12-km primary sampling units (PSU) was then selected randomly within each sampling region. The first-stage sample selection was designed to spread the sample geographically within each region. The target number of PSUs per region was 55. This was a subjective decision based on balancing the desire to disperse the sample spatially within each region while still retaining the cost advantage of clustering sample pixels within a limited number of PSUs. From a precision standpoint, if the sample would result in approximately one sample pixel per PSU for a given land-cover class, the upper bound (assuming true user's accuracy of 50%) on the estimated user's accuracy would be 7% for this sample size of 55 pixels. This was deemed an acceptable bound on precision. One PSU per frame cell was selected randomly resulting in all PSUs in a region having the same inclusion probability. If the number of PSUs selected in a region exceeded the target, the sample size of PSUs was reduced to the target by selecting an equal probability subsample from the initial draw of PSUs. If the initial draw returned fewer PSUs than the target, a second draw of PSUs from each frame cell was taken, with each PSU in the region again having an equal inclusion probability. Sampling was without replacement so that each PSU selected in the initial draw was not eligible for selection in the second draw. If the number of PSUs exceeded the target number after the second draw, an equal probability subsample of the total number of PSUs from both draws was selected to reduce the number of PSUs to the target. The subsample selection was independent of the frame cell stratification, so it could happen, for example, that both PSUs from a frame cell could be retained in the final sample, or that neither of the two PSUs selected were retained.

The third layer of stratification was the map land-cover class. In each of the 10 regions, 100 sample pixels of each class were selected via stratified random sampling from the first-stage sample PSUs selected in the region. A pixel was the secondary sampling unit (SSU) in the two-stage cluster design. The decision to use a pixel as the spatial unit of assessment is consistent with the “best practice” recommendations suggested by Strahler et al. (2006, p. 9). All land-cover classes except perennial ice/snow (Table 1) were used as strata. Thus, 15,000 samples in total were collected for the assessment (10

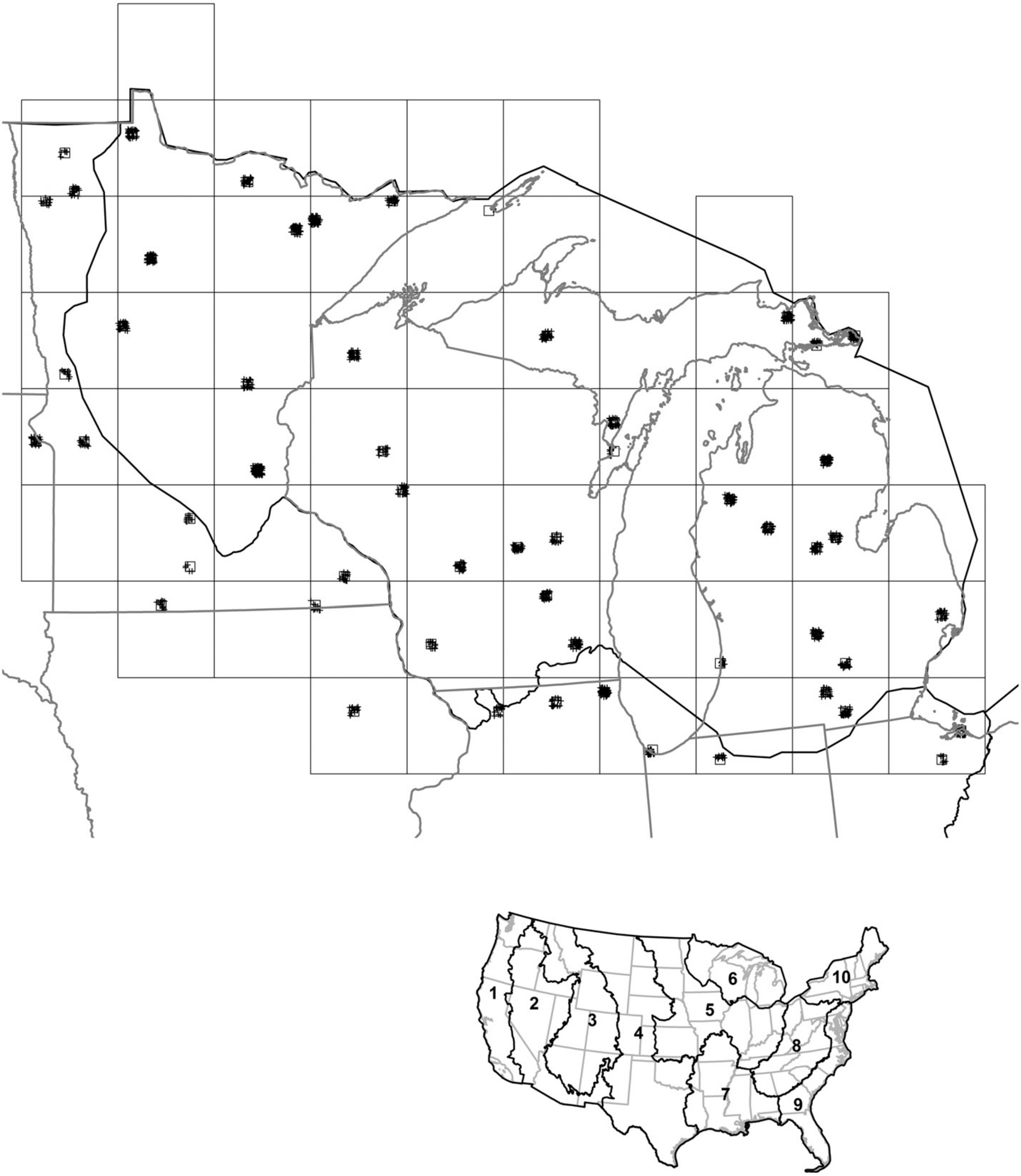

Fig. 1. Sampling design. The large squares represent the 120-km × 120-km frame cells and the smaller squares represent the selected 12-km × 12-km PSUs (all PSUs are not shown). The symbol used for sample pixels (crosshairs) sometimes obscures the boundaries of the selected PSUs. The black line depicts the region 6 boundary. The 120-km × 120-km cell boundaries were used to adjust the regional map boundaries so that all 120-km × 120-km cells and hence PSUs belonged to a single region (i.e., 120-km × 120-km cells and PSU cells were not split across regions). State boundaries are shown in gray. The inset map of the conterminous US shows the boundaries for all 10 regions (the geographic strata).

regions with 15 classes per region and generally 100 samples per class per region). Perennial ice/snow was not included because it was found in only 4 of the 10 regions, and comprised very small proportions of the area in these regions (Table S1). The additional cost of collecting reference data for this class was not justified given the rarity of perennial ice and snow.

For a few rare land-cover classes in selected regions, the sample pixels were selected without the constraint of the first-stage sample of PSUs. That is, the sample pixels were selected from all pixels mapped as that class in the region. This deviation from the standard sampling protocol was implemented to avoid having almost all sample pixels located within a single PSU when a class was very rare and highly concentrated spatially within a region. This sample selection protocol was used for classes 11, 23, and 24 in Region 2, classes 23 and 24 in Regions 3 and 5, and class 11 in Region 4.

The sampling design implemented for the NLCD 2001 assessment achieved two desirable design criteria typically sought for large-area accuracy assessments. Stratification by map land-cover class achieved the objective of precise class-specific estimates of accuracy, and clustering reduced the cost of the assessment. Combining both stratification and clustering can be done in many ways, and the advantages and disadvantages of different options are discussed by Stehman (2009).

2.2. Response design and definition of agreement

Reference land-cover classifications were obtained for each sample pixel by photointerpretation of Digital Orthophoto Quarter Quadrangles (DOQQ). These raster media have a nominal spatial resolution of 1 m^2. Reference sample locations were selected from the Albers equal area projection used for NLCD products and re-projected into the native UTM projections used for DOQQ products. Other available raster media (e.g., IKONOS) were used when DOQQs were not available. Four teams of interpreters, located throughout the conterminous United States, carried out the reference classification protocol. All reference data for a given region were collected by a single team (i.e., a given region was not split across teams).

The protocols for reference data collection included: 1) blind interpretation; 2) collection of primary and alternate reference labels; 3) assignment of a nominal level of confidence in the chosen reference label or labels; 4) inclusion of the date of the imagery used for determining the reference land-cover classes, and; 5) consistency in reference label assignment within and across teams. Interpreters were not provided *a priori* knowledge of the mapped land-cover class (i.e., "blind interpretation") to avoid interpreter bias in assigning reference class labels. The photointerpreters were allowed to assign an alternate land-cover label, in addition to the primary reference land-cover label, when they judged that more than one label was appropriate. Approximately 85% of the reference sample pixels included an alternate label. Each reference label was accompanied by a nominal self-assessment of photointerpreter confidence in the label. The nominal categories were "not confident," "somewhat confident," and "confident." A rating of "confident" was assigned to 77% of the reference samples, and a rating of "somewhat confident" was assigned to 21% of the reference samples. Photointerpreters rarely used the "not confident" rating (2%). The reference data also included the dates of reference imagery acquisition so that they could be compared with the map acquisition dates to determine if time lags in image acquisition were associated with classification errors (Congalton & Green, 1993; Wickham et al., 2004). Consistency in reference label assignment was enhanced in two ways. First, within each team, approximately 5% of the reference labels were checked by another member of the team to foster consistency in reference label assignments among team members. These checks were used to stimulate further review of potentially difficult cases and to establish commonality of approach when interpreting similar difficult cases. Second, bi-weekly, web-enabled conference calls among the teams were used to discuss sample points that presented interpretation issues. The web-enabled calls allowed members of all teams to simultaneously view sample points overlaid on the reference media and Landsat composite images. These points were discussed among members of all teams in order to reach consensus on reference label assignment. The web-enabled conference calls were included as a protocol to promote consistent reference label assignment within and among photointerpretation teams.

The final accuracy assessment dataset contained eight (8) primary attributes. Attributes derived from the reference data included: 1) the primary label, 2) the alternate label, 3) photointerpreter confidence, and 4) the acquisition date of reference media. Attributes derived from the map included: 5) the sample (center) pixel map label, 6) the modal map label(s) from the 3 × 3 pixel window surrounding the sample pixel, 7) image (i.e., Landsat) acquisition date, and 8) the number of different map land-cover classes in the 3 × 3 pixel window centered on the sample pixel (hereafter, heterogeneity). The full accuracy assessment dataset can be used to define agreement in several different ways, and to examine how agreement is affected by factors such as the time lag between map and reference image sources, spatial misregistration between map and reference labels, confidence in reference label assignment, and the spatial heterogeneity derived from the map land-cover classes surrounding the sample pixel. We briefly describe a few examples. The option to assign an alternate reference label was included because some land-cover class definitions are inherently fuzzy rather than crisp (Lunetta et al., 2001; Powell et al., 2004). Differences in map and reference labels can arise because of spatial misregistration between map and reference labels (Lanter & Veregin, 1992; Verbyla & Hammond, 1995), and positional errors for this assessment were assumed to occur because of the differences in spatial resolution of the 30-m × 30-m map versus the 1-m × 1-m reference data, the inherent geometric error of the map and reference media, and different geographic projections. Alternate reference labels can also be used to account for spatial misregistration between map and reference labels (Hagen, 2003). In heterogeneous areas, the photointerpreter can use a land-cover class adjacent to the sample pixel as the alternate reference label to account for the impact of misregistration on agreement. Likewise, defining agreement as a match between reference labels and map modal classes accounts for spatial misregistration between map and reference media when one or more modal map classes is different than the map label of the sample pixel. Comparison of agreement by heterogeneity can be used to determine edge effects on agreement (Smith et al., 2002, 2003; van Oort et al., 2004). When heterogeneity is equal to one, the sampled pixel is surrounded on all sides by like-classified pixels and is therefore not on a boundary (i.e., edge) between two land-cover classes. Heterogeneity is greater than one when the sampled pixel is on the edge between two or more land-cover classes.

The primary definitions of agreement used to report accuracy were: 1) the map land-cover class of the sample pixel matched either the primary or alternate reference land-cover label, and 2) a modal map land-cover class matched either the primary or alternate reference land-cover label. All modal map classes were considered for determining agreement when there was no majority in the 3 × 3 pixel window centered on the sample pixel. The first definition is called center agreement, and the second definition is called modal agreement.

The primary difference between the response design protocols of the two NLCD accuracy assessments was that a much more formal communication and coordination protocol was implemented among interpreter teams in the 2001 assessment to foster greater consistency among interpreters. The response design implemented for NLCD 1992 included the same attributes in its accuracy assessment data set and used similar definitions of agreement (Stehman et al., 2003; Wickham et al., 2004).

2.3. Analysis

The analysis is derived from the general estimation theory of probability sampling (cf. Särndal et al., 1992), which requires determining the inclusion probabilities resulting from the sampling protocol (Stehman & Czaplewski, 1998; Stehman, 2001). An inclusion probability is defined as the probability that a particular pixel is included in the sample. Inclusion probabilities are necessary to construct statistically consistent estimates of accuracy. The two-stage structure of the sampling design generates an inclusion probability for each stage. The first-stage inclusion probability, π_{1u}, is determined by the protocol used to select the sample of PSUs. By construction, all geographic strata within a mapping region had the same number of PSUs, K. Each pixel within a PSU was sampled with the same inclusion probability associated with the PSU within which the pixel was contained, so $\pi_{1u}=k/K$ for each pixel in the mapping region. At the second stage, those pixels selected in the first-stage sample were stratified by their mapped land-cover class. Suppose N_h^* pixels mapped as class h were selected in the first-stage sample of PSUs. A simple random sample of n_h pixels of map class h was selected from the N_h^* pixels available. Conditional on the selected first-stage sample, the second-stage inclusion probability for each pixel of class h was $\pi_{2.1hu}=n_h/N_h^*$. Consequently, the inclusion probability of pixel u, incorporating both stages of sampling (Särndal et al., 1992, Chapter 9), was

$$\pi_{hu} = \pi_{2.1hu}\pi_{1u} = \left(n_h / N_h^*\right)(k / K). \tag{1}$$

The inclusion probabilities are known for all pixels in the sample, and they are greater than zero for all pixels in the mapping region. These two conditions establish the probability sampling basis of the design. Eq. (1) also shows that within each mapping region, all pixels mapped as Level II land-cover class h have the same inclusion probability.

Stratified random sampling formulas were applied to estimate the error matrix and associated summary measures. We next develop these general estimation formulas. Let $y_{hu}(i, j)$ be the observation recorded for sample pixel u, where the h subscript indicates that pixel u was selected from stratum h. Define $y_{hu}(i, j)=1$ if the agreement definition results in pixel u belonging to map class i and reference class j in the error matrix; otherwise, $y_{hu}(i, j)=0$ (i.e., pixel u does not fall into cell (i, j) of the error matrix). Note that i and j may refer either to an Anderson Level I or Level II class, but h is always a Level II class determined by the original stratification by Level II map class. The value of $y_{hu}(i, j)$ depends on the definition of agreement employed. The estimation weight associated with pixel u is the reciprocal of the inclusion probability,

$$w_{hu} = 1 / \pi_{hu} = \left(KN_h^*\right) / (kn_h). \tag{2}$$

The weight, w_{hu}, is not affected by the definition of agreement because it is determined by the sampling design, not the response design.

Within each of the 10 geographic strata, the parameter N_{ij}, the number of pixels in the stratum that belong to cell (i, j) of the error matrix, is estimated by

$$\hat{N}_{ij} = \sum_{u \in s} w_{hu}\, y_{hu}(i, j) \tag{3}$$

where $\sum_{u \in s}$ indicates summation over all sample pixels, and the total number of pixels in a geographic stratum is estimated by

$$\hat{N} = \sum_{u \in s} w_{hu} \tag{4}$$

The cell proportions of the error matrix are then estimated by $\hat{p}_{ij}=\hat{N}_{ij}/N$. The estimators of overall, user's, and producer's accuracy (Story & Congalton, 1986) are as follows (q is the number of land-cover classes):

$$\text{Overall accuracy} = \sum_{i=1}^{q} \hat{p}_{ii}$$
$$\text{User's accuracy of map class } i = \hat{p}_{ii} / \sum_{j=1}^{q} \hat{p}_{ij}$$
$$\text{Producer's accuracy of reference class } j = \hat{p}_{jj} / \sum_{i=1}^{q} \hat{p}_{ij}.$$

The variance estimators follow the approach discussed in Stehman et al. (2003, Sec. 5.2), with one exception. For the NLCD 1992 variance estimators, a map polygon was treated as the "cluster" and pixels within the same map polygon were treated as secondary sampling units within that cluster. The variance estimators used in this NLCD 2001 assessment treat the 12-km × 12-km PSU as the cluster. Because the number of map polygons is expected to exceed the number of PSUs, variances computed on the basis of the PSUs as the clusters would be anticipated to be slightly higher than variances computed using a map polygon as the cluster. Computations were conducted using the Statistical Analysis Software (SAS 2003, Version 9.1.3, SAS Institute, Inc., Cary, North Carolina, USA).

3. Results

The 10 regional error matrices are reported in the online Supplementary Material (Tables S3–S12). Based on the mode definition of agreement, the nationwide, overall thematic accuracies are 78.7% at Level II and 85.3% Level I (these modal accuracies and other national averages reported herein are unweighted averages of the 10 regional estimates). The nationwide, overall thematic accuracies of 78.7% (Level II) and 85.3% (Level I) for NLCD 2001 are approximately 20% and 5% higher than the corresponding accuracy statistics for NLCD 1992 (Fig. 2). At the class-specific level, there are improvements in NLCD 2001 cropland (Level II) and forest (Level I) accuracies relative to NLCD 1992 accuracies. Nationwide, NLCD 2001 cropland user's accuracy (Table 2) and producer's accuracy (Table 3) average 82% and 88%, respectively, whereas NLCD 1992 cropland user's and producers accuracies average 43% and 54%, respectively. The improvement in overall agreement is also reflected in the 87.0% average class-specific user's accuracy for forest (Table 2), compared to 80% forest user's accuracy for NLCD 1992. The national NLCD 2001 forest user's accuracy improves to 91.5% when region 2, which is dominated by shrubland, is excluded. The national NLCD 2001 and NLCD 1992 forest producer's accuracies are 88.5% and 86.1%, respectively. The national Level II and Level I class-specific accuracies are adversely affected by two regions (7 and 9) with noticeably poorer results. Overall accuracies in these two regions are about 10% lower at Level II and 5% lower at Level I than the other eight regions.

Aggregating Level II classes to Level I improves overall accuracy from 78.7% at Level II to 85.3% at Level I (Table 2). This suggests a significant portion of the misclassification cuts across the NLCD 2001 classification hierarchy (e.g., class 21 misclassified as 81). The most noticeable occurrence of cross-hierarchy classification error occurs in region 2. The region is approximately 70% shrubland, and although the user's accuracy for shrubland in this region is high (82.8%), misclassification with grassland is present. Shrubland–grassland misclassification is not resolved by aggregation within the classification hierarchy, resulting in only a 2% improvement in overall accuracy through aggregation from Level II to Level I. Forest and urban are exceptions to the pattern of significant cross-hierarchy misclassification. User's accuracies for the Level II forest (deciduous, evergreen, mixed) and urban (open space, and low, medium, and high intensity development) classes are generally much lower than the overall regional user's accuracies. However, forest and urban user's accuracies improve by approximately 20% when aggregated to Level I, indicating that a substantial portion of the misclassification is among Level II classes that were nested within the Level I forest and urban classes.

Use of the modal rather than the center pixel map label generally improves user's accuracies by 1% to 2% (Table 2). In contrast, the NLCD 1992 accuracy assessment reported 15%–20% improvements in user's accuracy when using the modal map label compared to using the center pixel map label (Stehman et al., 2003; Wickham et al., 2004). The decision to use a five-pixel minimum mapping unit (mmu) for NLCD 2001 (Homer et al., 2007) probably accounts for the smaller percentage gain in user's accuracy for the modal-based map

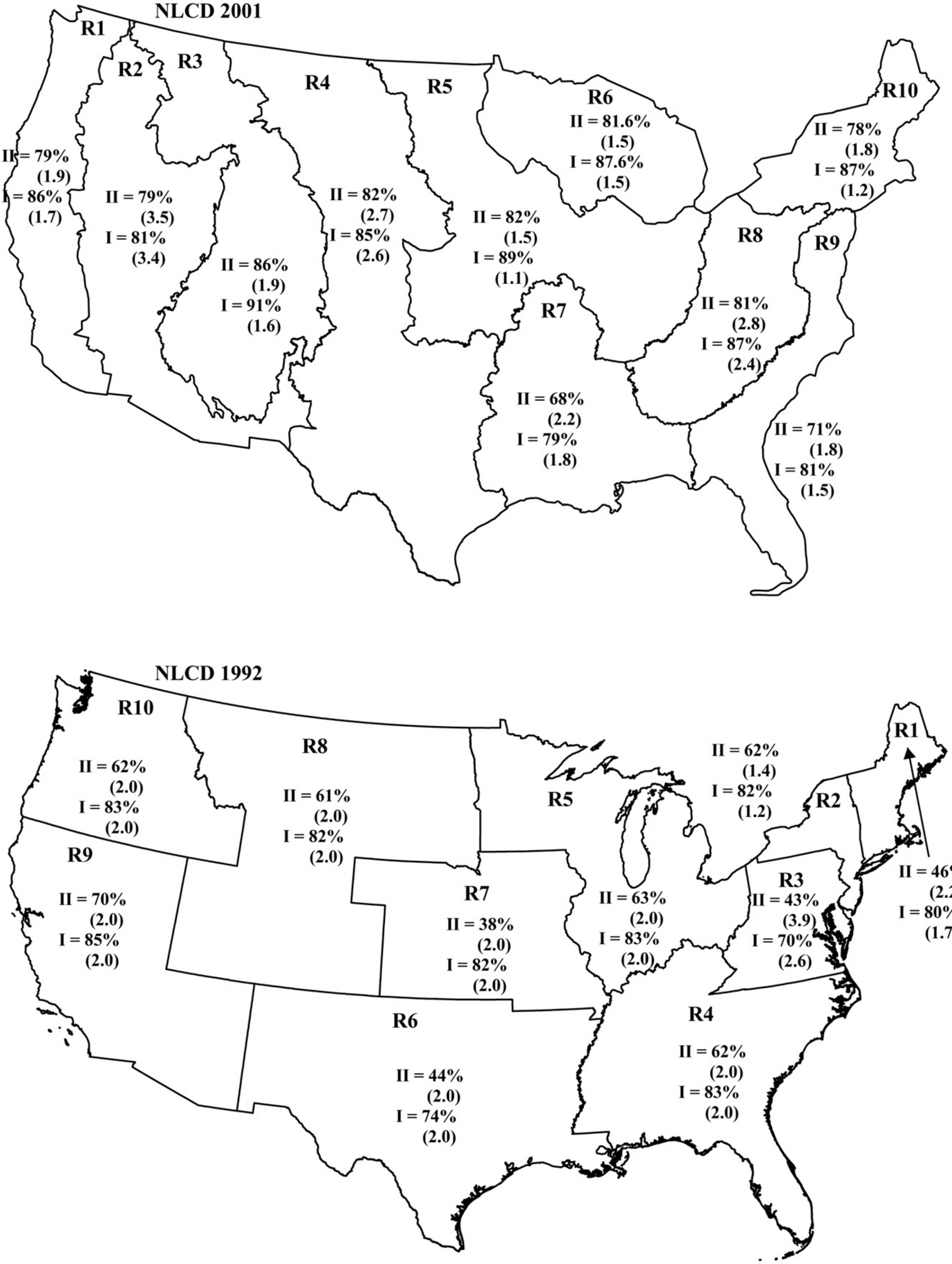

Fig. 2. Regional overall accuracies for NLCD 2001 (top) and NLCD 1992 (bottom) based on the mode definition of agreement. Overall accuracies are rounded to the nearest whole percentage. Standard errors for the overall accuracies are in parentheses. The labels "Rx" identify the regions used to geographically stratify the sample (e.g., R1 = region 1). NLCD 1992 accuracy results were reported by EPA administrative regions.

Table 2
Regional user's accuracies for Level II (top) and Level I (bottom). The row labeled I vs. II is the improvement in overall accuracy realized by aggregating the map classes from Level II to Level I.

Class	User's accuracy																			
	Region 1		Region 2		Region 3		Region 4		Region 5		Region 6		Region 7		Region 8		Region 9		Region 10	
	Center	Mode	Center	Mode	Center	Mode	Center	Mode	Center	Mode	Center	Mode	Center	Mode	Center	Mode	Center	Mode	Center	Mode
11	94.0	92.6	94.7	92.4	97.0	96.9	90.0	91.2	86.0	90.0	98.0	94.9	94.0	95.7	94.0	94.7	94.2	98.6	94.0	93.6
21	40.0	44.3	37.0	39.0	49.0	47.2	25.0	17.4	42.0	43.5	83.0	81.1	55.0	71.8	78.0	84.5	82.0	79.8	74.0	77.5
22	52.0	51.5	21.0	29.6	42.0	34.8	73.0	77.5	60.0	67.0	87.0	78.4	38.0	40.8	88.0	82.6	76.0	77.2	76.0	72.2
23	76.0	79.6	72.0	65.8	76.0	71.7	58.0	58.9	56.0	56.4	88.0	86.4	46.0	43.9	85.4	88.7	67.0	63.3	71.0	73.6
24	97.0	98.2	86.7	90.1	70.7	76.0	58.0	66.2	58.0	55.9	83.0	84.1	61.0	63.0	89.1	81.7	95.0	94.3	86.0	86.7
31	51.0	56.8	91.0	96.1	82.0	71.1	53.0	57.0	65.0	75.6	42.0	53.0	36.0	32.2	35.0	41.5	16.0	17.3	47.0	27.1
41	27.0	28.0	6.0	5.8	60.0	62.2	64.0	66.1	81.0	81.9	83.0	78.7	79.0	79.7	84.0	82.8	67.0	66.6	85.0	82.5
42	90.0	91.7	48.0	50.1	79.0	80.2	76.0	71.8	37.0	43.0	92.0	91.0	71.0	68.7	88.0	89.5	84.0	83.8	90.0	87.6
43	48.0	53.4	2.0	4.1	68.0	73.0	70.0	90.1	56.0	67.5	80.0	75.2	71.0	75.2	80.0	88.0	80.0	84.2	89.0	88.4
52	71.0	71.6	83.0	82.8	92.0	93.6	89.0	87.2	26.0	26.4	67.0	62.8	54.0	53.1	36.0	32.7	58.0	64.0	54.0	56.8
71	82.0	84.9	100.0	99.9	92.0	90.0	83.0	82.0	69.0	70.3	54.0	46.2	61.0	58.5	16.0	15.0	25.0	29.1	33.0	35.7
81	54.0	51.6	51.0	52.8	76.0	77.4	64.0	47.6	84.0	84.5	76.0	76.7	79.0	78.8	82.0	87.5	61.0	58.3	73.0	73.9
82	88.0	87.3	88.0	87.0	65.0	60.0	92.0	91.5	90.0	89.4	89.0	87.2	80.0	79.0	77.0	79.6	80.0	84.5	75.0	72.6
90	18.0	20.1	47.0	47.0	37.0	48.9	48.0	55.2	7.0	7.2	83.0	79.7	14.0	11.8	58.0	53.9	57.0	56.2	37.0	41.9
95	32.0	34.6	91.0	94.4	32.0	37.0	55.0	56.7	57.0	63.8	60.0	62.2	47.0	48.9	46.0	57.7	42.0	52.1	44.0	46.5
Overall	76.0	78.5	78.2	78.8	85.5	86.4	83.3	82.1	80.1	81.8	84.1	81.6	68.0	68.3	80.2	81.4	69.7	70.6	78.2	77.7

Class	User's accuracy																			
	Region 1		Region 2		Region 3		Region 4		Region 5		Region 6		Region 7		Region 8		Region 9		Region 10	
	Center	Mode	Center	Mode	Center	Mode	Center	Mode	Center	Mode	Center	Mode	Center	Mode	Center	Mode	Center	Mode	Center	Mode
10	94.0	94.3	94.7	92.8	97.0	78.8	90.0	91.2	86.0	90.0	98.0	95.0	94.0	95.7	94.0	93.7	99.2	98.6	94.0	94.5
20	74.3	82.9	60.4	70.0	69.0	76.8	70.4	57.0	66.6	70.7	91.7	92.9	68.5	81.6	85.5	89.7	93.1	92.9	84.7	87.0
30	51.0	58.1	91.0	96.1	82.0	71.7	53.0	56.3	65.0	76.2	43.3	56.2	36.0	32.4	35.0	43.2	16.0	19.4	47.0	51.4
40	95.7	96.1	46.6	47.5	88.9	90.6	89.7	87.0	88.0	87.9	90.9	89.7	89.4	89.8	91.6	90.4	96.5	96.0	96.5	95.8
50	71.0	72.1	83.0	82.8	92.0	93.7	89.0	87.2	26.0	27.5	67.0	62.6	54.0	52.6	36.0	34.4	58.0	62.4	54.0	57.7
70	82.0	84.9	100.0	99.9	92.0	90.8	83.0	82.0	69.0	70.2	54.0	46.8	61.0	58.0	16.0	15.9	25.0	28.6	33.0	35.4
80	82.5	82.3	83.5	84.6	78.1	77.1	94.5	94.6	97.3	96.7	90.8	89.3	89.9	89.4	84.6	88.7	75.9	77.4	80.4	80.9
90	26.7	30.5	84.8	87.1	44.5	51.9	63.4	67.8	39.2	44.9	84.5	83.4	24.1	25.2	66.0	63.5	57.0	56.5	42.7	48.0
Overall	84.1	86.1	80.0	80.5	89.6	90.6	86.3	85.2	88.0	89.1	89.0	87.6	77.9	78.9	86.0	86.8	79.2	80.7	86.1	87.4
I vs II	8.1	7.6	1.8	1.7	4.1	4.2	3.0	3.1	7.9	7.3	4.9	6.0	9.9	10.6	5.8	5.4	9.5	10.1	7.9	9.7

Table 3
Regional producer's accuracies for Level II (top) and Level I (bottom).

Level II	Producer's accuracy																			
	Region 1		Region 2		Region 3		Region 4		Region 5		Region 6		Region 7		Region 8		Region 9		Region 10	
Class	Center	Mode	Center	Mode	Center	Mode	Center	Mode	Center	Mode	Center	Mode	Center	Mode	Center	Mode	Center	Mode	Center	Mode
11	88.4	87.0	99.5	93.4	86.7	87.5	78.3	76.5	86.3	86.8	95.5	96.6	82.5	80.6	60.1	65.6	70.6	72.5	86.2	86.3
21	34.7	28.8	39.1	28.0	49.1	34.3	83.8	64.0	68.0	57.3	53.8	45.4	40.2	29.4	68.4	67.5	45.2	43.2	45.7	33.9
22	45.5	40.1	51.1	44.7	28.6	26.9	27.7	35.9	59.3	70.8	80.1	71.4	70.1	72.7	69.5	66.1	69.5	69.9	81.4	73.4
23	82.5	81.5	84.6	75.2	45.0	34.4	79.3	75.2	54.2	46.0	90.5	80.6	46.7	49.5	76.6	78.4	73.7	68.9	71.4	69.3
24	53.5	59.2	6.0	6.5	20.0	17.2	79.5	76.6	41.8	36.5	98.5	97.4	73.7	68.2	84.1	74.9	50.0	51.1	58.6	54.4
31	63.3	62.1	99.8	98.7	51.6	47.8	19.7	18.5	18.3	18.8	56.2	58.9	14.3	11.1	51.2	49.5	23.8	23.8	60.2	59.5
41	13.6	9.2	91.8	91.8	70.8	69.8	39.0	31.9	80.4	81.7	85.8	84.3	60.0	60.6	92.2	93.1	75.2	76.9	89.3	90.7
42	84.9	89.4	96.8	98.2	93.4	95.4	93.9	90.8	36.0	40.5	82.5	74.3	84.8	84.4	72.3	77.6	82.4	82.9	71.9	70.3
43	62.4	68.0	2.3	2.3	7.1	4.9	6.5	9.3	10.9	8.5	70.1	57.6	60.1	62.2	60.7	65.1	37.3	38.4	83.7	82.1
52	84.1	86.8	96.6	96.8	89.3	90.2	89.4	87.7	5.3	2.6	38.8	33.3	50.6	50.7	46.5	37.1	26.5	27.6	65.6	70.1
71	75.3	78.6	24.2	24.9	89.3	91.0	94.9	94.0	84.6	85.3	56.3	52.4	38.7	38.4	100.0	100.0	53.9	60.6	35.8	31.0
81	64.2	65.0	82.0	81.3	63.3	64.5	16.8	15.8	65.1	66.3	72.9	71.5	73.4	75.1	67.8	69.5	82.1	84.5	77.8	79.9
82	85.4	87.1	92.4	94.2	92.2	93.6	87.7	86.1	91.6	94.2	95.8	96.7	89.4	92.1	95.1	94.5	86.4	89.4	78.6	80.2
90	52.7	49.0	100.0	100.0	63.9	62.4	49.5	46.7	71.6	68.4	89.8	82.2	92.2	87.8	85.0	84.5	92.5	89.5	83.8	87.0
95	34.5	34.4	83.2	77.0	56.4	55.2	33.2	32.2	76.8	80.2	72.6	65.1	84.2	76.3	17.3	10.2	74.2	79.3	31.0	29.2

Class	Producer's accuracy																			
	Region 1		Region 2		Region 3		Region 4		Region 5		Region 6		Region 7		Region 8		Region 9		Region 10	
	Center	Mode	Center	Mode	Center	Mode	Center	Mode	Center	Mode	Center	Mode	Center	Mode	Center	Mode	Center	Mode	Center	Mode
10	88.4	87.0	99.7	93.4	87.9	85.1	78.7	76.5	87.4	88.3	95.5	96.6	88.8	88.0	60.1	65.4	78.0	74.2	88.6	87.7
20	72.5	70.0	50.3	43.0	56.5	56.3	97.9	87.0	80.6	76.8	73.3	69.7	64.0	61.5	74.3	74.3	66.3	67.6	68.2	62.0
30	65.1	63.9	99.9	98.8	51.6	48.1	22.2	20.7	20.6	20.5	86.8	85.9	15.2	11.4	51.2	47.1	50.0	50.0	65.7	64.8
40	87.8	90.6	99.0	99.6	92.8	94.4	83.3	79.4	82.0	81.9	90.3	88.7	77.1	77.6	93.3	94.2	84.8	85.8	90.4	92.5
50	86.4	89.1	97.2	97.4	92.8	93.7	90.9	89.1	5.4	2.6	39.9	36.5	54.1	57.4	60.6	50.5	31.2	29.8	78.8	83.9
70	76.5	79.3	24.6	25.0	89.8	91.1	95.4	94.3	88.9	89.0	67.7	62.8	42.7	42.0	100.0	100.0	57.4	62.0	35.8	30.2
80	86.0	87.2	96.3	97.1	75.3	76.0	73.6	74.8	91.7	93.5	94.1	94.4	89.7	91.6	73.9	74.9	87.7	90.4	83.7	85.2
90	50.5	46.9	96.3	92.1	71.3	68.9	47.3	45.7	80.3	82.2	94.1	87.4	95.2	90.9	86.7	84.3	94.4	94.3	75.7	77.0

agreement definition. The 5-pixel mmu protocol substantially increases the odds that the map class of the sample (center) pixel is also a mode class.

A geographic pattern in classification error related to class rarity is evident from the accuracy results. Shrubland and grassland user's accuracies decrease from west to east (Table 2). These classes are abundant in the west (regions 1 through 4) but generally rare in the east (regions 5 through 10). Conversely, deciduous forest user's accuracy decreased from east to west, and this too is correlated with the proportion of deciduous forest in the sampling regions. The positive relationship between class abundance and accuracy is also a pattern observed in the NLCD 1992 accuracy assessment (Stehman et al., 2003; Wickham et al., 2004) and in other mapping studies (e.g., Foody, 2005; Thompson & Gergel, 2008).

The regional error matrices (Supplementary Material) reveal three other error patterns. First, the context of grass is difficult to distinguish. Misclassification among developed open space, grassland, pasture, and cropland, which are all defined by grass, is 3.5% in the west and 4.4% in the east. Second, developed open space (class 21) producer's accuracies tend to be lower than user's accuracies due to omission errors with abundant classes. The disparity between producer's and user's accuracies for developed open space indicates that the class tends to "look like its surroundings." The pattern is more apparent in the eastern US (Supplementary Material, regions 5 through 10) because of the notably higher percentages of urban. Third, producer's accuracies for woody wetlands are much higher than their user's accuracies, principally because reference labels for woody wetland sample pixels are commonly one of the 3 upland forest classes. It is apparently difficult for the map makers, the reference photointerpreters, or both to distinguish "wet" from "dry" forest, and it is impossible to determine from the available data if one of the two sources (map, reference) is a more significant contributor to the misclassification.

The response design implemented permits estimating accuracy by various subsets of the sample to determine how different aspects of reference data and map context affect accuracy results (Table 4). Including an alternate label in the definition of agreement has the most substantial impact. Defining agreement as a match between the map label and either the primary or alternate reference label improves overall accuracy by approximately 20% at both levels of the classification hierarchy relative to defining agreement as a match between the map label and only the primary reference label. The user's and producer's accuracies, by region, based on using only the primary reference label are documented in the Supplementary Material (Table S2). Photointerpreter confidence in reference label assignment and heterogeneity (i.e., number of map classes in the 3×3 window surrounding the sampled pixel) also affect map accuracy. Level II overall accuracy improves by approximately 3.5% when only the subset of reference samples with a rating of "confident" is used. Similarly, overall accuracy improves by approximately 7% using the subset that is not on the edge between two or more land-cover classes. However, this subset of homogeneous area represents only about one-third of the total sample. A significantly higher error rate for "edge" pixels was also reported for NLCD 1992 (Smith et al., 2002, 2003; Stehman et al., 2003; Wickham et al., 2004). As noted above, choice of center versus mode definition of agreement has little effect on overall accuracy.

Time lags between reference and map image acquisition dates have little effect on agreement. Based on a logistic regression model, the probability of agreement is not significantly associated with the difference between reference and map image acquisition dates. A similar result was observed for the NLCD 1992 accuracy assessment (Wickham et al., 2004). Time lags between reference and map image sources are intuitively regarded as a potential source of disagreement because of the possibility of land-cover change occurring during the interval between acquisitions of map and reference sources (Congalton & Green, 1993). Land-cover change is rare (Biging et al., 1999; Fry et al., 2009), and samples for reference data acquisition are also rare. Land-cover change and sampling are independent events, suggesting that the spatial pattern of each would have to overlay in a very unlikely manner for land-cover change to strongly influence overall or class-specific accuracies. Rather than time, there is anecdotal evidence that the imagery used to collect the reference data influenced agreement. Imagery available through Google Earth was the reference source for approximately 125 samples in region 4 due to unavailability of other reference media. Agreement for this admittedly small subset is about 15% lower than the overall accuracy for the region.

4. Discussion

The conterminous national NLCD 2001 Level II and Level I thematic user's accuracies are approximately 20% and 5% higher than the corresponding statistics for NLCD 1992. The NLCD 1992 accuracy assessment results (Stehman et al., 2003; Wickham et al., 2004) contributed to the changes in mapping methods used for NLCD 2001

Table 4
Regional overall accuracies by different definitions of agreement. The agreement definitions "Center" and "Mode" are defined in the Methods. "Center Pri Only" and "Mode Pri Only" are the counterparts of "Center" and "Mode," but include only the primary reference label for determining agreement. "High Conf" refers to those samples whose nominal confidence rating in the reference label assignment was "confident" (see Methods). "Homogeneous" refers to the subset of sample pixels whose 3×3 pixel neighborhood included only like-classified pixels. Agreement for "High Conf" and "Homogeneous" is defined based on a match with either primary or alternate reference label. The "pri only" results are conspicuously low for region 2 because the region was strongly dominated by class 52 (Shrub/Scrub) and there was a strong tendency for the photointerpreters to assign class 71 (Grassland/Herbaceous) as the primary label and class 52 as the alternate label to the sample pixels for class 52.

Agreement def.	Region 1	Region 2	Region 3	Region 4	Region 5	Region 6	Region 7	Region 8	Region 9	Region 10	Regional average
Level II											
Center	76.0	78.2	85.5	83.3	80.1	84.1	68.0	80.2	69.7	78.2	78.3
Mode	78.5	78.8	86.4	82.1	81.8	81.6	68.3	81.4	70.6	77.7	78.7
Center Pri Only	56.5	18.5	53.4	57.1	63.5	68.5	52.5	63.9	53.9	54.6	54.2
Mode Pri Only	58.3	18.4	54.7	55.9	64.8	66.2	52.3	63.3	53.1	54.7	54.2
High Conf	81.0	80.0	91.0	85.7	87.1	85.1	71.1	85.5	75.2	81.3	82.3
Homogeneous	70.0	82.0	90.7	89.0	88.5	89.3	79.0	94.9	84.2	88.6	85.6
Level I											
Center	84.1	80.0	89.6	86.3	88.0	89.0	77.9	86.0	79.2	86.1	84.6
Mode	86.1	80.5	90.6	85.2	89.1	87.6	78.9	86.8	80.7	87.4	85.3
Center Pri Only	69.8	21.1	56.9	61.5	79.0	68.5	68.2	79.6	69.5	79.3	65.3
Mode Pri Only	71.0	20.9	58.2	60.5	80.6	78.7	68.3	79.6	80.7	80.6	67.9
High Conf	87.2	81.4	93.2	87.7	92.7	89.2	81.1	89.5	84.1	89.3	87.5
Homogeneous	93.5	80.4	95.0	90.8	95.3	93.8	84.8	96.4	88.9	93.2	91.2

(Homer et al., 2004), and these methodological changes appear to have had a positive effect on data quality. It is likely that the improved discrimination of cropland and forest will expand the NLCD user-community. For example, dasymetric approaches to assignment of pesticide application rates to cropland from county-level statistics can be used to assess more confidently the impact of pesticides on aquatic resources. The improved forest user's accuracies provide better data for an already broad user community (e.g., Riitters et al., 2004; Heinz Center, 2008).

The design used for the NLCD 2001 land-cover accuracy assessment conforms to and advances many of the accepted protocols for land-cover thematic accuracy assessment (Foody, 2002). Reference data were collected using a probability-based sampling design, thereby permitting rigorous statistical inference (e.g., statistically consistent estimators of overall and class-specific accuracy and estimation of standard errors) (Stehman, 2001). The sampling design included stratification to avoid small sample sizes for rare land-cover classes (Zhu et al., 2000) and to account for geographic variation in accuracy. The response design incorporated protocols to foster consistent assignment of reference labels, thereby diminishing some of the impact of interpreter variability observed by Mann and Rothley (2006) in their study in which interpreters were allowed to work independently. It also included alternate reference labels and modal map values, which in turn were used to construct different definitions of agreement. Such 'scaling' of agreement can be used to account for disagreement between map and reference labels due to locational error (Lanter & Veregin, 1992; Verbyla & Hammond, 1995; Hagen, 2003) and inherent fuzziness in class definitions (Lunetta et al., 2001; Powell et al., 2004). Inclusion of the variety (number) of land-cover classes in a 3 × 3 pixel window surrounding the sample can be used to examine agreement in relation to land-cover class boundaries (Wickham et al., 1997; Smith et al., 2002, 2003), and use of a photointerpreter confidence rating can be used to gauge the effect of reference data quality on agreement. The lessons learned from research on land-cover accuracy assessment reveal that agreement is not a binary concept (Congalton & Green, 1999; Khorram et al., 1999; Foody, 2002; Mann & Rothley, 2006). A variety of factors affect agreement and reporting a range of agreement scores better accounts for these factors. Our dataset can be used to examine most of the factors that are known to affect agreement.

Summarizing Congalton (1994), Foody (2002) recounts the history of thematic accuracy assessment from qualitative visual inspections to the present standard of comparison of reference and map classifications that are reported using error matrices. Because of the now well established use of reference data, reference data quality is a recurrent topic in thematic accuracy assessment (Foody, 2009). Recognizing that reference data are not error free, these discussions generally conclude that reported thematic map accuracies can be biased by poor reference data quality (Powell et al., 2004; Foody, 2009), and that higher reference data quality would remove that bias, resulting in higher thematic map accuracies (Congalton & Green, 1999; Khorram et al., 1999; Foody, 2002; Mann & Rothley, 2006). The response designs implemented in two NLCD accuracy assessments included protocols to account for reference data error. These analyses (e.g., Table 4), while useful, cannot be used to adjust the accuracy estimates or to reduce the standard errors to account for reference data quality. Use of auxiliary data through double sampling (Stehman, 1996) is one approach to thematic map accuracy assessment that accounts for reference data quality. In the case of NLCD, ground visits for, say, 30 sample pixels per land-cover class in each region could be used to construct the second phase of a double or two-phase sampling design, in which the reference data from the ground visits could be viewed as adjusting the accuracy estimates derived from the first-phase sample from DOQQs. The use of double sampling increases costs, but it is also likely that other evaluations of reference data quality would also increase costs. Given the widespread acceptance of reference data as a means of assessing land-cover thematic accuracy and the data quality issues that surround their use (Foody, 2002), it seems logical that future research should evaluate the use of auxiliary data as a more quantitatively rigorous means of reference data quality assessment.

Acknowledgements

This manuscript has been subjected to USGS peer review, and it has been approved for publication. M. Coffey, R. Compton, C. Costello, J. Dewitz, B. Fiehler, L. Gass, T. Hale, G. Henson, T. Meunier, R. Poss, C. Rich, D. Sorenson, and D. Wheeler collected the reference data.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.rse.2010.01.018.

References

Biging, G. S., Colby, D. R., & Congalton, R. G. (1999). Sampling systems for change detection accuracy assessment. In R. S. Lunetta, & C. D. Elvidge (Eds.), *Remote sensing change detection: Environmental monitoring methods and applications* (pp. 281—308). London: Taylor and Francis.

Carr, M. H., Hoctor, T. D., Goodison, C., Zwick, P. D., Green, J., Hernandez, P., et al. (2002). *Final Report: Southeastern Ecological Framework.* URL: http://www.geoplan.ufl.edu/epa/index.html (accessed January 21, 2010).

Cohen, W. B., & Goward, S. N. (2004). Landsat's role in ecological applications of remote sensing. *BioScience*, *54*, 535—545.

Congalton, R. G. (1994). Accuracy assessment of remotely sensed data: Future needs and directions. *Proceedings of Pecora 12 Land Information from Space-based Systems* (pp. 383—388). Bethesda: American Society for Photogrammetry and Remote Sensing (ASPRS).

Congalton, R. G., & Green, K. (1993). A practical look at sources of confusion in error matrix generation. *Photogrammetric Engineering and Remote Sensing*, *59*, 641—644.

Congalton, R. G., & Green, K. (1999). *Assessing the accuracy of remotely sensed data: Principles and practices.* Boca Raton: Lewis Publishers.

Doherty, J., & Johnston, J. M. (2003). Methodologies for calibration and predictive analysis of a watershed model. *Journal of the American Water Resources Association (JAWRA)*, *39*, 251—265.

Foody, G. M. (2002). Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, *80*, 185—201.

Foody, G. M. (2005). Local characterization of thematic classification accuracy through spatially constrained confusion matrices. *International Journal of Remote Sensing*, *26*, 1217—1228.

Foody, G. M. (2009). The impact of imperfect ground reference data on the accuracy of land cover change estimation. *International Journal of Remote Sensing*, *30*, 3275—3281.

Fry, J. A., Coan, M. J., Homer, C. G., Meyer, D. K., & Wickham, J. D. (2009). *Completion of the National Land Cover Database (NLCD) 1992–2001 land cover change retrofit product: U.S. Geological Survey Open-File Report 2008-1379.* 18 pp. URL: http://pubs.usgs.gov/of/2008/1379 (accessed January 21, 2010).

Hagen, A. (2003). Fuzzy set approach to assessing similarity in categorical maps. *International Journal of Geographical Information Science*, *17*, 235—249.

Heilman, G. E., Jr., Strittholt, J. R., Slosser, N. C., & 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 (2008). The State of the Nation's Ecosystems 2008: Measuring the Lands, Waters, and Living Resources of the United States. *The H. John Heinz III Center for Science, Economics, and the Environment*. Washington, DC: Island Press.

Hoekstra, J. M., Boucher, T. M., Ricketts, T. H., & Roberts, C. (2005). Confronting a biome crisis: Global disparities of habitat loss and protection. *Ecology Letters*, *8*, 23—29.

Homer, C. G., DeWitz, J., Coan, M., Hossain, N., Larson, C., Herold, N., et al. (2007). Completion of the 2001 National Land Cover Database for the conterminous United States. *Photogrammetric Engineering and Remote Sensing*, *73*, 337—341.

Homer, C., Huang, C., Yang, L., Wylie, B., & Coan, M. (2004). Development of a 2001 National Landcover Database for the United States. *Photogrammetric Engineering and Remote Sensing*, *70*, 829—840.

Jackson, L. E., Hillborn, E. D., & Thomas, J. C. (2006). Towards landscape design guidelines for reducing Lyme disease risk. *International Journal of Epidemiology*, *35*, 315—322.

Khorram, S., Biging, G. S., Chrisman, N. R., Colby, D. R., Congalton, R. G., Dobson, J. E., et al. (1999). Accuracy assessment of remote sensing-derived change detection. *Monograph*,Bethesda: American Society for Photogrammetry and Remote Sensing (ASPRS) 64 pp.

Lanter, D. P., & Veregin, H. (1992). A research paradigm for propagating error in layer-base GIS. *Photogrammetric Engineering and Remote Sensing*, *58*, 825—833.

Lunetta, R. S., Iiames, J., Knight, J., Congalton, R. G., & Mace, T. H. (2001). An assessment of reference data variability using a "virtual field reference database". *Photogrammetric Engineering and Remote Sensing*, *63*, 707—715.

Mann, S., & Rothley, K. D. (2006). Sensitivity of Landsat/IKONOS accuracy comparison to errors in photointerpreted reference data and variations in test point sets. *International Journal of Remote Sensing*, *27*, 5027−5036.

Marshall, C. H., Pielke, R. A., Sr., Steyaert, L. T., & 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.

Milesi, C., Elvidge, C. D., Nemani, R. R., & Running, S. W. (2003). Assessing the impact of urban development on net primary productivity in the southeastern United States. *Remote Sensing of Environment*, *86*, 401−410.

Nolan, B. T., Hitt, K. J., & Ruddy, B. C. (2002). Probability of nitrate contamination of recently recharged groundwaters in the conterminous United States. *Environmental Science and Technology*, *36*, 2138−2145.

Powell, R. L., Matzke, N., de Souza, C., Jr., Clark, M., Numata, I., Hess, L. L., et al. (2004). Sources of error in accuracy assessment of thematic land-cover maps in the Brazilian Amazon. *Remote Sensing of Environment*, *90*, 221−234.

Radeloff, V. C., Hammer, R. B., Stewart, S. I., Fied, J. S., Holcomb, S. S., & McKeefry, J. F. (2005). The wildland–urban interface in the United States. *Ecological Applications*, *15*, 799−805.

Riitters, K. H., Wickham, J. D., & Coulston, J. W. (2004). A preliminary assessment of the Montréal process indicators of forest fragmentation for the United States. *Environmental Monitoring and Assessment*, *91*, 257−276.

Riitters, K. H., Wickham, J. D., O'Neill, R. V., Jones, K. B., Smith, E. R., Coulston, J. W., et al. (2002). Fragmentation of continental United States forests. *Ecosystems*, *5*, 815−822.

Särndal, C. E., Swensson, B., & Wretman, J. (1992). *Model-assisted survey sampling.* New York: Springer-Verlag.

Smith, J. H., Wickham, J. D., Stehman, S. V., & Yang, L. (2002). Impacts of patch size and land cover heterogeneity on thematic image classification accuracy. *Photogrammetric Engineering and Remote Sensing*, *68*, 65−70.

Smith, J. H., Stehman, S. V., Wickham, J. D., & Yang, L. (2003). Effects of landscape characteristics on land-cover class accuracy. *Remote Sensing of Environment*, *84*, 342−349.

Stehman, S. V. (1996). Use of auxiliary data to improve the precision of estimators of thematic map accuracy. *Remote Sensing of Environment*, *58*, 169−176.

Stehman, S. V. (2001). Statistical rigor and practical utility in thematic map accuracy assessment. *Photogrammetric Engineering and Remote Sensing*, *67*, 727−734.

Stehman, S. V. (2009). Sampling designs for accuracy assessment of land cover. *International Journal of Remote Sensing, 30,* 5243–5272.

Stehman, S. V., & Czaplewski, R. L. (1998). Design and analysis for thematic map accuracy assessment. *Remote Sensing of Environment*, *64*, 331−344.

Stehman, S. V., Wickham, J. D., Smith, J. H., & 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.

Stehman, S. V., Wickham, J. D., Wade, T. G., & Smith, J. H. (2008). Designing a multi-objective, multi-support accuracy assessment of the 2001 National Land Cover Data (NLCD 2001) of the United States. *Photogrammetric Engineering and Remote Sensing*, *74*, 1561−1571.

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

Strahler, A. H., Boschetti, L., Foody, G. M., Friedl, M. A., Hansen, M. C., Herold, M., et al. (2006). *Global land cover validation: Recommendations for evaluation and accuracy assessment of global land cover maps, EUR 22156 EN-DG.*Luxembourg: Office for Official Publications of the European Communities 48 pp.

Thompson, S. D., & Gergel, S. E. (2008). Conservation implications of mapping rare ecosystems using high spatial resolution imagery: Recommendations for heterogeneous and fragmented landscapes. *Landscape Ecology*, *23*, 1023−1037.

van Oort, P. A. J., Bregt, A. K., de Bruin, S., & de Wit, A. J. W. (2004). Spatial variability in classification accuracy of agricultural crops in the Dutch national land-cover database. *International Journal of Geographical Information Science*, *18*, 611−626.

Verbyla, D. L., & Hammond, T. O. (1995). Conservative bias in classification accuracy assessment due to pixel-by-pixel comparison of classified images with reference grids. *International Journal of Remote Sensing*, *16*, 581−587.

Vogelmann, J. E., Howard, S. M., Yang, L., Larson, C. R., Wylie, B. K., & Van Driel, J. N. (2001). Completion of the 1990's National Land Cover Data Set for the conterminous United States. *Photogrammetric Engineering and Remote Sensing*, *67*, 650−652.

Wagner, R. C., Dillaha, T. A., & Yagow, G. (2007). An assessment of the reference watershed approach for TMDLs with biological impairments. *Water, Air, & Soil Pollution*, *181*, 341−354.

Weber, T. (2004). Landscape ecological assessment of the Chesapeake Bay watershed. *Environmental Monitoring and Assessment*, *94*, 39−53.

Weber, T., Sloan, A., & Wolf, J. (2006). Maryland's green infrastructure assessment: Development of a comprehensive approach to land conservation. *Landscape and Urban Planning*, *77*, 94−110.

Wickham, J. D., O'Neill, R. V., Riitters, K. H., Wade, T. G., & Jones, K. B. (1997). Sensitivity of landscape metrics to land cover misclassification and differences in land cover composition. *Photogrammetric Engineering and Remote Sensing*, *63*, 397−402.

Wickham, J. D., Stehman, S. V., Smith, J. H., & Yang, L. (2004). Thematic accuracy of MRLC-NLCD land cover for the western United States. *Remote Sensing of Environment*, *91*, 452−468.

Zhu, Z., Yang, L., Stehman, S. V., & Czaplewski, R. L. (2000). Accuracy assessment of the U.S. Geological Survey regional land-cover mapping program: New York and New Jersey region. *Photogrammetric Engineering and Remote Sensing*, *66*, 1425−1435.