Contents lists available at SciVerse ScienceDirect

Remote Sensing of Environment

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

Accuracy assessment of NLCD 2006 land cover and impervious surface

James D. Wickham [a,*], Stephen V. Stehman [b], Leila Gass [c], Jon Dewitz [d], Joyce A. Fry [d], Timothy G. Wade [a]

[a] U.S. EPA, Office of Research and Development, National Exposure Research Laboratory, MD: E243-05, Research Triangle Park, NC 27711, USA
[b] SUNY College of Environmental Science and Forestry, 320 Bray Hall, 1 Forestry Dr., Syracuse, NY 13210, USA
[c] U.S. Geological Survey, 520N. Park Ave., Tucson, AZ 87519, USA
[d] U.S. Geological Survey, EROS Data Center, 47914 252nd St., Sioux Falls, SD 57198, USA

ARTICLE INFO

Article history:
Received 20 August 2012
Received in revised form 30 November 2012
Accepted 1 December 2012
Available online 4 January 2013

Keywords:
Impervious surface accuracy assessment
Forest loss
Land-cover change accuracy assessment
MRLC
Stratified sampling
Urbanization
United States

ABSTRACT

Release of NLCD 2006 provides the first wall-to-wall land-cover change database for the conterminous United States from Landsat Thematic Mapper (TM) data. Accuracy assessment of NLCD 2006 focused on four primary products: 2001 land cover, 2006 land cover, land-cover change between 2001 and 2006, and impervious surface change between 2001 and 2006. The accuracy assessment was conducted by selecting a stratified random sample of pixels with the reference classification interpreted from multi-temporal high resolution digital imagery. The NLCD Level II (16 classes) overall accuracies for the 2001 and 2006 land cover were 79% and 78%, respectively, with Level II user's accuracies exceeding 80% for water, high density urban, all upland forest classes, shrubland, and cropland for both dates. Level I (8 classes) accuracies were 85% for NLCD 2001 and 84% for NLCD 2006. The high overall and user's accuracies for the individual dates translated into high user's accuracies for the 2001–2006 change reporting themes water gain and loss, forest loss, urban gain, and the no-change reporting themes for water, urban, forest, and agriculture. The main factor limiting higher accuracies for the change reporting themes appeared to be difficulty in distinguishing the context of grass. We discuss the need for more research on land-cover change accuracy assessment.

Published by Elsevier Inc.

1. Introduction

Each release of the MultiResolution Land Characteristics (MRLC) consortium National Land Cover Database (NLCD) has represented an advance in Landsat-based land-cover mapping. NLCD 1992 (Vogelmann et al., 2001) was the first 30 m×30 m land-cover product for the continental United States. NLCD 2001 (Homer et al., 2007) built upon the success of NLCD 1992 by incorporating Alaska, Hawaii, and Puerto Rico into the product, by introducing a database concept to land-cover mapping through inclusion of percentage urban impervious surface and percentage forest canopy cover, and by improving the methods used for land-cover classification. NLCD 2006 in turn has built upon NLCD 2001 by incorporating land-cover and impervious surface change for the continental United States (Fry et al., 2011; Xian et al., 2011).

With the production and release of NLCD 2006, MRLC has initiated a shift from land-cover mapping to land-cover monitoring (Fry et al., 2011). Whereas NLCD 1992 and NLCD 2001 were land-cover databases, NLCD 2006 is a land-cover change database that includes land cover and impervious surface for 2001 and 2006 and change between the two target years. Using NLCD 2001 as the baseline, spectral change was evaluated to identify areas of land-cover and impervious surface change that were then re-classified to produce NLCD 2006 (Fry et al., 2011; Xian et al., 2011, 2009). Pixels that were not identified as having undergone spectral change retained their NLCD 2001 land-cover label. The spectral change analysis and re-classification were conducted on a path/row by path/row basis using a single, leaf-on scene from each target year (2001 and 2006) that were acquired on near anniversary dates (within two weeks of each other). Land-cover change applies to the full 16-class (i.e., Level II) legend in NLCD 2006 rather than the simplified 8-class (i.e., Level I) legend (Table 1).

Accuracy assessment is an established protocol of the NLCD mapping process. Accuracy assessments have followed the completion of the NLCD 1992 (Stehman et al., 2003; Wickham et al., 2004) and NLCD 2001 efforts (Wickham et al., 2010). The NLCD land-cover accuracy assessments have been supported by research on the impact of spatial pattern on agreement (J.H. Smith et al., 2002, 2003), how to prioritize among many accuracy assessment objectives in the face of cost constraints (Stehman et al., 2008), how to scale the information in error matrices to estimate the accuracy of land-cover proportions over much larger areas (counties, watersheds) (Stehman et al., 2009), and the role of the spatial unit used to assess agreement (Stehman & Wickham, 2011). In this paper, we report accuracy statistics for the NLCD 2006 product, including both land-cover change and impervious surface change. The accuracy assessment objectives include: 1) accuracy of both the 2001 and 2006 land-cover maps for the 16-class Level II and 8-class Level I classification hierarchies, 2) accuracy of change versus no change and accuracy of priority Level I change classes, and 3) accuracy of impervious surface change. These

[*] Corresponding author at: U.S. Environmental Protection Agency, Mail Drop: E243-05, 109 T.W. Alexander Dr., Research Triangle Park, NC 27711, USA. Tel.: +1 919 541 3077; fax: +1 919 541 4329.
E-mail address: wickham.james@epa.gov (J.D. Wickham).

0034-4257/$ – see front matter. Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.rse.2012.12.001

Table 1
NLCD Level II class codes and definitions. Level I classes are grouped by the tens' digit of the Level II class codes, e.g., Level I class code 40 (upland forest) includes Level II codes 41, 42, and 43. Definitions are from http://www.mrlc.gov/nlcd01_leg.php. Classes found only in Alaska are not included in the table.

Class code	Description
11	Open water—areas of open water, generally with less than 25% cover of vegetation or soil.
12	Perennial ice/snow—areas characterized by a perennial cover of ice and/or snow, generally greater than 25% of total cover.
21	Developed, open space—areas with a mixture of constructed materials, but mostly vegetation in the form of lawn grasses. Impervious surfaces account for less than 20% of total cover. The areas most commonly include large-lot single-family housing units, parks, golf courses, and vegetation planted in developed settings for recreation, erosion control or esthetic purposes.
22	Developed, low intensity—areas with a mixture of constructed materials and vegetation. Impervious surfaces account for 20% to 49% of total cover. These areas commonly include single-family housing units.
23	Developed, medium intensity—areas with a mixture of constructed materials and vegetation. Impervious surfaces account for 50% to 79% of total cover. These areas commonly include single-family housing units.
24	Developed, high intensity—highly developed areas where people reside or work in large numbers. Examples include apartment complexes, row houses, and commercial/industrial. Impervious surfaces account for 80% to 100% of total cover.
31	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 cover. More than 75% of tree species shed foliage simultaneously in response to seasonal changes.
42	Evergreen forest—areas dominated by trees generally greater than 5 meters tall, and greater than 20% of total cover. More than 75% of tree species maintain their leaves all year. Canopy is never without green foliage.
43	Deciduous forest—areas dominated by trees generally greater than 5 meters tall, and greater than 20% of total cover. Deciduous or evergreen species are not 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 shrubs, young trees in an early successional stage or trees stunted from environmental conditions.
71	Grassland/herbaceous—areas dominated by gramanoid or herbaceous vegetation, generally greater than 80% of total vegetation cover. 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 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 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 20% of vegetative cover and the soil or substrate is periodically saturated with or covered with water.

three main objectives support users' needs for data quality information for static and dynamic investigations that require land-cover data.

2. Methods

Accuracy assessment protocols for NLCD 2006 generally follow those established for the 1992 and 2001 assessments (Stehman et al., 2003; Wickham et al., 2004, 2010). We briefly review the NLCD 1992 and NLCD 2001 protocols focusing on the sampling design, response design, and analysis components of accuracy assessment (Stehman & Czaplewski, 1998). The sampling design for the NLCD 1992 and 2001 accuracy assessments used a two-stage cluster sample with three levels of stratification. The main elements of the response design for the NLCD 1992 and 2001 accuracy assessments were: 1) use of a pixel as the spatial unit for the accuracy assessment; 2) blind collection (i.e., interpreters have no a priori knowledge of map land cover labels) of primary and alternate reference labels for the sample pixels, and; 3) nominal confidence rating for the reference land-cover labels. The analysis component of the NLCD 1992, 2001 and 2006 accuracy assessments is based on general estimation theory of probability sampling (cf. Särndal et al., 1992), which requires determination of the inclusion probabilities resulting from the sampling protocol (Stehman, 2001; Stehman & Czaplewski, 1998). The main difference between the NLCD 1992 and 2001 accuracy assessments sampling designs and the NLCD 2006 sampling design is that cluster sampling was used in 1992 and 2001 but not in 2006. For both the 1992 and 2001 assessments cluster sampling was used because of the cost and convenience of spatially constraining the sample pixels to a limited number of clusters. For the NLCD 1992 and 2001 assessments, Digital Orthophoto Quarter Quadrangles (DOQQs) were the primary source of reference data imagery with other high-resolution aerial photographic sources also used on occasion (Wickham et al., 2004, 2010). Cluster sampling was used to reduce the number of DOQQs that had to be purchased and handled. For the NLCD 1992 assessment, these DOQQs were in hard copy (paper) form and had to be purchased and handled manually. Although greater online access to DOQQs improved for the NLCD 2001 assessment, the convenience and cost savings of using fewer DOQQs allowed by a cluster sampling design were still warranted. For the NLCD 2006 accuracy assessment, online access to reference imagery was greatly enhanced relative to the NLCD 1992 and 2001 assessments and the savings or convenience of controlling the spatial distribution of the sample pixels to a limited number of clusters was no longer a compelling reason to implement a cluster sampling design. In general, standard errors of accuracy estimates would be expected to be smaller for a design that does not incorporate clusters (Gallego, 2012; Stehman, 1997; Wickham et al., 2004). Although the current and previous NLCD accuracy assessments have employed sampling designs focusing on a single spatial assessment unit (30 m pixel), we acknowledge the possibility of a multi-scale approach that would permit assessing accuracy at several spatial scales (cf. Pontius & Cheuk, 2006; Stehman, 2009; Stehman & Wickham, 2011; Olofsson et al., 2012; Stehman et al., 2012).

Accuracy assessment of land-cover change was not an objective of the previous NLCD assessments. Assessment of land-cover change accuracy introduces trade-offs between cost and precision (Biging et al., 1998). The geometric increase in the number of classes makes it infeasible (from a cost perspective) to collect a sufficient sample size to estimate class-specific accuracies with acceptable levels of precision for all classes. For example, for the 8 Level I land-cover classes there are 64 possible classes for the 2001 to 2006 change product, the 8 no-change or “stable” classes and the 56 possible transitions. Rather than attempt to assess all classes, the problem can be addressed by choosing a set of classes that will be the primary focus of the accuracy assessment. MRLC members (http://www.mrlc.gov) were surveyed to identify the priority classes for the NLCD 2006 assessment. Considering the full NLCD legend (Table 1), MRLC members recommended a balanced assessment focusing on the priority change and no-change classes listed in Table 2. The sampling design was then constructed to

Table 2
NLCD 2006 strata (A) and reporting themes (B). The column # is the number of samples pixels per region. The strata were used to select the sample and the reporting themes were used to report accuracy results. The reporting themes are not mutually exclusive.

A

Strata	# of samples per region	Description
Water loss	50	From water (2001) to any other class (2006)
Water gain	50	From any class (2001) to water (2006)
Forest to urban	75	From any upland forest class (2001) to any urban class (2006)
S/G to urban	75	From shrubland (S) or grassland (G) (2001) to any urban class (2006)
Agric. to urban	75	From cropland or pasture (2001) to any urban class (2006)
Wetland to urban	50	From woody or herbaceous wetland (2001) to any urban class (2006)
Forest to S/G	75	From any upland forest class (2001) to shrubland (S) or grassland (G) (2006)
F/S/G to agric.	75	From any upland forest (F) class, shrubland (S) or grassland (G) (2001) to cropland or pasture (2006)
Agric. to forest	50	From cropland or pasture (2001) to any upland forest class (2006)
S/G flux	75	From shrubland (S)(2001) to grassland (G) (2006), and vice versa
Agric. to S/G	50	From cropland or pasture (2001) to shrubland (S) or grassland (G) (2006)
S/G to forest	75	From any upland forest class (2001) to shrubland (S) or grassland (G) (2006)
Wetland-agric. flux	75	From woody or herbaceous wetland (2001) to cropland or pasture (2006), and vice versa
Water, no change	25	Classified as water on both dates, not including perennial ice and snow
Urban, no change	75	Classified as any urban class on both dates
Barren, no change	25	Classified as barren on both dates
Forest, no change	75	Classified as any upland forest class on both dates
Shrub., no change	75	Classified as shrubland on both dates
Grass., no change	50	Classified as grassland on both dates
Agric., no change	75	Classified as any agriculture class both dates
Wetland, no change	50	Classified as any wetland class on both dates.
Other	200	A “catch-all” stratum that included all possible from-to changes not identified above.

B

Reporting themes	Description
Water loss	From water (2001) to any other class (2006)
Water gain	From any class (2001) to water (2006)
Urban gain	From any non-urban class (2001) to any urban class (2006)
Forest loss	From any upland forest class (2001) to any non-upland forest class (2006)
Forest gain	From any non-upland forest class (2001) any upland forest class (2006)
Shrubland loss	From shrubland (2001) to any other class but shrubland (2006)
Shrubland gain	From any class but shrubland (2001) to shrubland (2006)
Grassland loss	From grassland (2001) to any other class but grassland (2006)
Grassland gain	From any class but grassland (2001) to grassland (2006)
Water, no change	Classified as water on both dates, not including perennial ice and snow
Urban, no change	Classified as any urban class on both dates
Forest, no change	Classified as any upland forest class on both dates
Shrub., no change	Classified as shrubland on both dates
Grass., no change	Classified as grassland on both dates
Agric., no change	Classified as any agriculture class both dates

allocate the limited accuracy assessment resources (i.e., sample size) in a manner to enhance the precision of the accuracy estimates of the priority classes identified. Specifically, class-level stratification (Table 2) was implemented within each of the 10 geographic regions corresponding to NLCD 2001 mapping zones (Fig. 1). The regional stratification was implemented to capture geographic differences in class-specific accuracies, and we retained the geographic regions of the original NLCD 2001 accuracy assessment (Wickham et al., 2010) to facilitate comparisons to the accuracy results of the previous NLCD 2001 land-cover product (Homer et al., 2007). Incorporating class-level stratification within a regional stratification was a feature of the sampling designs used for previous NLCD accuracy assessments (Stehman et al., 2003; Wickham et al., 2010). Class-level stratification is motivated by the objective of estimating user's accuracy for targeted classes. In particular, classes (whether a land-cover class or a land-cover change class) that are rare may not be represented by a large enough sample size to yield precise estimates of user's accuracy so defining these classes as strata allows for increasing the sample size from these classes. The target sample size of 1500 pixels per region was deemed affordable and consistent with the sample size of the accuracy assessment conducted for the original NLCD 2001 (Wickham et al., 2010).

The spatial support unit for the accuracy assessment was a pixel. Although there is debate over the most appropriate spatial support unit for accuracy assessment, use of pixels for accuracy assessment is a well-established practice that avoids many of the complications that arise when other spatial support units (e.g., blocks of pixels, polygons) are used (Stehman & Wickham, 2011). Sample pixels were initially located on Landsat imagery and then co-located on high resolution imagery. The primary sources of high resolution imagery were 1 m–2 m resolution data from the National Agricultural Imagery Program (NAIP), provided by the United States Department of Agriculture (USDA) National Agriculture Statistics Service (NASS), and Google Earth. Reference land-cover labels were based on interpretation of the high resolution imagery. Color composites of Landsat images served as the reference media when NAIP and Google Earth imagery were not available for the sample pixel. Lack of NAIP and Google Earth imagery occurred for about 10% of the sample pixels in 2001 and 3% of the sample pixels in 2006. The relatively new Google Earth slider time toolbar was helpful in determining whether or not change had taken place. The slider time toolbar allowed interpreters to view multiple dates of imagery for the same location.

Interpreters collected information for several attributes for each sampled pixel (Table 3), and as per standard protocol, interpreters were not informed of the map class of the sample pixels. The main attributes were the primary and alternate labels for each date (2001 and 2006). Primary labels were deemed the most likely land-cover label, and alternate labels were deemed the second most likely land-cover label. Use of primary

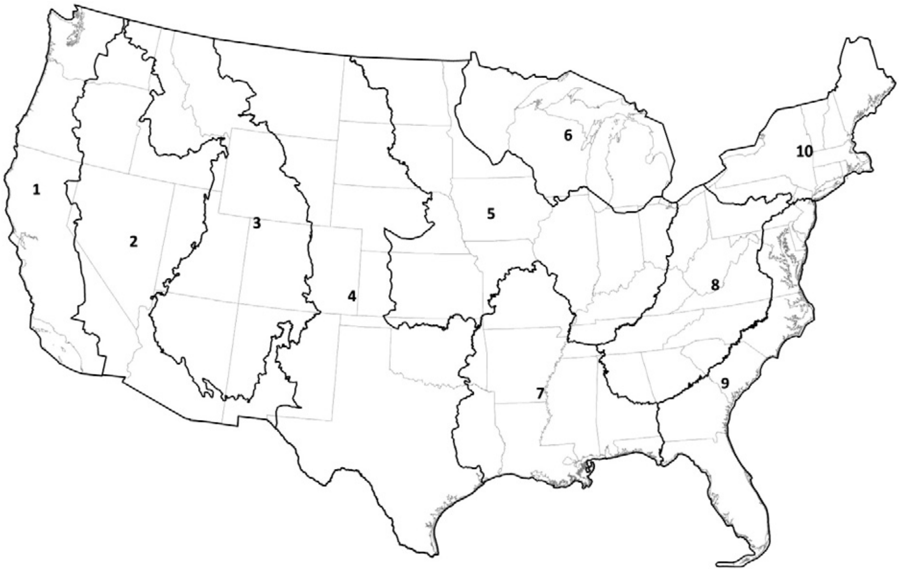

Fig. 1. Regional stratification for NLCD 2006 accuracy assessment.

and alternate land-cover labels to define agreement can be viewed as a special case of Gopal and Woodcock's (1994) linguistic scale, fuzzy approach to accuracy assessment (Stehman et al., 2003, p. 513). For each sample pixel, the interpreter also assigned a confidence value of 1 (not confident), 2 (somewhat confident), or 3 (confident), and these confidence values were used to determine if agreement was associated with interpreter confidence.

Land-cover change reference labels were generated by combining the reference labels for the individual dates. For example, a primary reference label of shrubland for 2001 and a primary reference label of grassland for 2006 would constitute a land-cover change. Four possible land-cover change labels arise from the combinations of primary and alternate labels for each date (primary–primary, primary–alternate, alternate–primary, and alternate–alternate). Change interpretation also included a binary assessment of whether change was apparent in the reference imagery. The binary interpretation was applied to four different attributes—no change, change within 0.45 ha (5 pixels), change within 1.08 ha (12 pixels), and change within 2.88 ha (32 pixels) (Fig. 2). The binary change attributes were used to match the mapping protocols for NLCD 2006 that were established to minimize single-pixel changes (Fry et al., 2011). Change to urban had to encompass an area of at least 0.45 ha, change to agriculture had to encompass an area of at least 2.88 ha, and change involving all other classes had to encompass an

Table 3
Reference dataset attributes.

Index	Description
1	Primary land-cover label, 2001
2	Alternate land-cover label, 2001
3	Nominal confidence, 2001 (1 = not confident; 2 = somewhat confident; 3 = confident)
4	Primary land-cover label, 2006
5	Alternate land-cover label, 2006
6	Nominal confidence, 2006 (1 = not confident; 2 = somewhat confident; 3 = confident)
7	No change (binary, 1 = no change occurred, 0 = change occurred)
8	Change at 0.45 ha (binary, 1 = change of at least 0.45 ha occurred, 0 = no change)
9	Change at 1.08 ha (binary, 1 = change of at least 1.08 ha occurred, 0 = no change)
10	Change at 2.88 ha (binary, 1 = change of at least 2.88 ha occurred, 0 = no change)
11	Date of reference media for 2001 (where available)
12	Date of Landsat data for 2001
13	Date of reference media for 2006 (where available)
14	Date of Landsat data for 2006
15	Change land-cover label 1, (primary 2001 vs. primary 2006)
16	Change land-cover label 2, (primary 2001 vs. alternate 2006)
17	Change land-cover label 3, (alternate 2001 vs. primary 2006)
18	Change land-cover label 4, (alternate 2001 vs. alternate 2006)

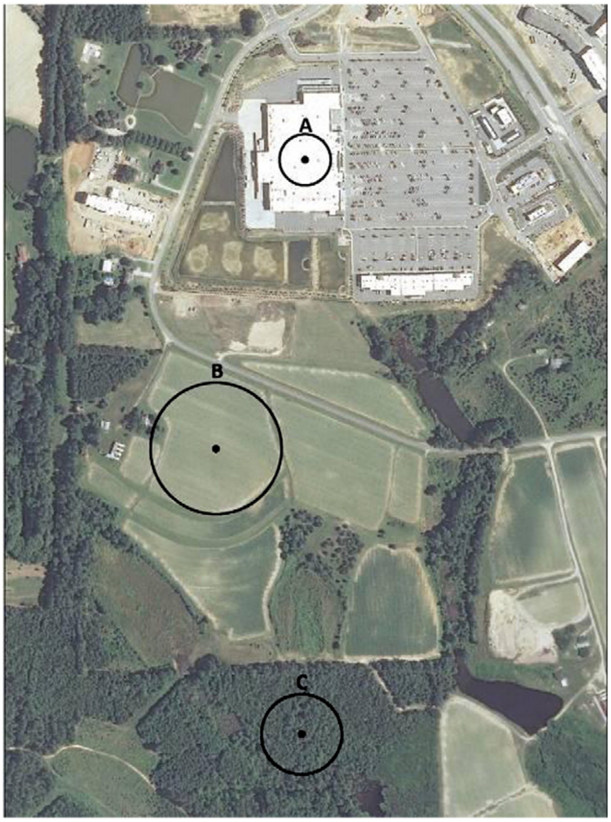

Fig. 2. Hypothetical sample pixels (points) in the context of NLCD 2006 minimum mapping unit thresholds for land-cover change. Change had to encompass at least 0.45 ha (A) for urban, at least 2.88 ha for agriculture (B), and at least 1.08 ha (C) for all other land-cover classes. Background imagery is 2010 NAIP photography of Wake County, North Carolina. Source—http://datagateway.nrcs.usda.gov.

area of at least 1.08 ha. For the no-change attribute, a value of one (1) indicated that change did not occur, whereas values of one (1) for the 0.45 ha, 1.08 ha, 2.88 ha change attributes indicated that change had occurred that encompassed 0.45 ha, 1.08 ha, or 2.88 ha, respectively. It was possible for interpreters to record a value of one (1) for the no change attribute and also record a value of one (1) for one or more of the three binary change attributes, which served as an indicator of uncertainty of land-cover change. In total, the reference dataset included 18 attributes (Table 3).

Four teams of interpreters at the U.S. Geological Survey collected the reference data. The interpreters participated in weekly web-enabled conference calls in which interpreters could share their assigned sample pixels and associated imagery with all other members of the interpretation teams. These web-enabled conference calls were used to promote consistent interpretation among individuals. In addition, a pilot study was conducted prior to the initiation of the accuracy assessment project. In the pilot, the interpretation teams "calibrated" the reference classification process by working with a common set of sites in northern Florida and central Colorado.

The sampling design was a stratified random sample. The 22 strata based on the map classification of a pixel are listed in Table 2. These 22 strata were defined within each of 10 regions to create 220 strata nationally (10 regions by 22 strata). The regional stratification was implemented for the objective of reporting accuracy on a regional basis, and the 22-class stratification was used to address the objective of estimating class-specific accuracy for the priority change and no change classes. The same general estimation formulas can be applied to produce regional estimates or national estimates simply by changing the number of strata involved in the estimator. For overall accuracy, the estimator is

$$\hat{O} = \left(\frac{1}{N}\right)\sum_{h=1}^{H} N_h \hat{p}_h$$

where $\hat{p}_h$ is the sample proportion of pixels correctly classified in stratum h, N_h is the number of pixels in stratum h, N is the total number of pixels in the region (or nation), and the summation is over all H strata ($H=22$ for a regional estimate and $H=220$ for a national estimate). Overall accuracy is estimated for both land cover products (2001 and 2006) and for land-cover change. Two versions of overall accuracy are computed for 2001–2006 land-cover change. In one case, overall accuracy is based on a binary map of change and no change (without regard to specific types of change and no change), and in the second case overall accuracy is computed for 2001–2006 taking into account the specific types of change or no change.

A ratio estimator is used to estimate user's and producer's accuracies because the general form of the parameter of interest for user's and producer's accuracies can be expressed as a ratio R = Y/X, where Y is the population total of y_u (i.e., from a census) where

$$y_u = \begin{cases} 1 & \textit{if pixel u satisfies condition A} \\ 0 & \textit{if pixel u does not satisfy condition A} \end{cases}$$

and X is the population total of x_u, where

$$x_u = \begin{cases} 1 & \textit{if pixel u satisfies condition B} \\ 0 & \textit{if pixel u does not satisfy condition B.} \end{cases}$$

For example, to estimate user's accuracy of class 21 (developed open space), condition A would be that pixel u is mapped as class 21 and that the reference class label is also class 21 (i.e., the pixel is classified correctly), and condition B would be that pixel u is mapped as class 21. The parameter for user's accuracy of class 21 would be the total number of pixels meeting condition A divided by the total number of pixels meeting condition B (both totals would be for the entire region, and would represent census values). This parameter can be estimated from the reference sample. Similarly, to estimate producer's accuracy of class 21, condition A would remain the same, but condition B would be that pixel u has reference class 21. The parameter for producer's accuracy of class 21 would be the total number of pixels satisfying condition A divided by the total number of pixels satisfying condition B. The combined ratio estimator (Cochran, 1977 Section 6.11,) for user's accuracy or producer's accuracy is then

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

where $\bar{x}_h$ is the sample mean of x_u in stratum h and $\bar{y}_h$ is the sample mean of y_u in stratum h.

The estimated variance of the combined ratio estimator is

$$\hat{V}\left(\hat{R}\right) = \left(\frac{1}{\hat{X}^2}\right)\left[\sum_{h=1}^{H} N_h^2 (1-n_h/N_h)\left(s_{yh}^2 + \hat{R}^2 s_{xh}^2 - 2\hat{R}s_{xyh}\right)/n_h\right]$$

where s_{yh}^2 and s_{xh}^2 are the sample variances of y_u and x_u for stratum h and s_{xyh} is the sample covariance of x_u and y_u in stratum h. Because the strata do not necessarily correspond to the land-cover types or change types for which user's and producer's accuracies may be of interest, contributions to the estimator may arise from more than one stratum. For example, for estimating user's accuracy of deciduous forest in 2001, several strata may contribute sample pixels mapped as deciduous forest in 2001: forest to urban; forest to shrub or grassland; forest, shrub, or grassland to agriculture; forest no change; and other. For a stratum in which no sampled pixels satisfy condition A (the condition defining the numerator of $\hat{R}$), $y_u=0$ for all sample pixels and $\bar{y}_h=0$ and $s_{yh}^2=0$. Similarly, for a stratum in which no sampled pixels satisfy condition B (the condition defining the denominator of $\hat{R}$), $x_u=0$ for all sample pixels and $\bar{x}_h=0$ and $s_{xh}^2=0$. These estimates were obtained using SAS (version 9.2, SAS Institute Incorporated, Cary, North Carolina, USA).

The ideal assessment of impervious surface change accuracy would require estimation of percentage urban impervious surface for each sample pixel from reference data (e.g., NAIP imagery). The cost to obtain these ideal reference data was prohibitive, so we resorted to a more limited assessment that took advantage of the reference land-cover data obtained to assess the accuracy of the NLCD 2006 land-cover products. The impervious surface change classification (Xian et al., 2011) was the basis for "to urban" changes between NLCD 2001 and NLCD 2006 (Fry et al., 2011). That is, if the impervious surface classification indicated an increase in impervious surface and the NLCD 2001 map label was not urban (Level I, class 20), the pixel was mapped as a change to urban. To assess the accuracy of the mapped impervious change, the reference data were collapsed to two groups, change to urban and not a change to urban. If a pixel mapped as having an increase in impervious surface also had a reference label indicating a change to urban, the pixel would be considered mapped correctly.

3. Results

Defining agreement as a match between the map label and either the primary or alternate reference label, Level II overall accuracies were 79% for NLCD 2001 (Table 4) and 78% for NLCD 2006 (Table 5) for the continental United States. User's accuracies were greater than 80% for water (class 11), high intensity developed (class 24), all upland forest classes (classes 41, 42, 43), shrubland (class 52), and the cropland class (82) for both dates. A primary area of disagreement occurred among the NLCD grass-dominated classes: developed open space (class 21), grassland (class 71), pasture/hay (class 81), cropland (class 82), and emergent wetland (class 95). All are dominated by herbaceous vegetation set in different contexts. Confusion among these classes accounted for approximately 26% of the classification error

Table 4
NLCD 2001 Level II error matrix. Cell entries are expressed as percent of area, and based on agreement defined as a match between the map class and either the primary or alternate reference class. User's accuracy (UA) and producer's accuracy (PA) are reported with standard errors (SE) in parentheses. Overall accuracy is 79% (0.8%). Main diagonal entries (agreement) are in bold.

Map	Reference																
	11	12	21	22	23	24	31	41	42	43	52	71	81	82	90	95	UA (SE)
11	**1.998**		0.016				0.001	0.007			0.008	0.019	0.008	0.029	0.050	0.004	93 (1)
12		**0.006**					0.001					0.004					54(18)
21	0.005		**1.554**	0.201	0.028		0.002	0.272	0.147	0.099	0.141	0.155	0.221	0.204	0.029	0.011	51(3)
22			0.307	**0.872**	0.091	0.016	0.014	0.016	0.018		0.041	0.019	0.038	0.054	0.016	0.018	57(4)
23	0.006		0.105	0.052	**0.491**	0.029	0.002		0.005		0.009	0.022				0.009	67(4)
24	0.006		0.009	0.009	0.005	**0.200**											87(6)
31	0.039		0.013	0.001		0.001	**0.831**	0.014	0.041	0.002	0.115	0.129	0.012	0.004	0.005	0.005	69(4)
41	0.001		0.146	0.034				**10.154**	0.839	0.291	0.252	0.096	0.105	0.088	0.196	0.001	83(2)
42			0.125					0.835	**10.534**	0.285	0.543	0.157	0.027	0.069	0.001		84(2)
43	0.001		0.018					0.082	0.072	**1.669**	0.041	0.060			0.024		85(5)
52	0.008		0.354	0.019			0.095	0.595	0.614	0.293	**18.517**	0.580	0.388	0.217	0.009	0.157	85(2)
71	0.007		0.342	0.037			0.001	0.287	0.196	0.029	0.372	**11.055**	1.457	0.714	0.001	0.118	76(3)
81			0.292	0.016			0.017	0.189	0.033	0.002	0.202	0.253	**5.554**	0.106	0.007	0.039	83(2)
82	0.040		0.227	0.009	0.007		0.001	0.101	0.018	0.001	0.224	0.190	0.896	**14.247**	0.055	0.017	89(2)
90	0.011		0.016		0.030			1.160	0.362	0.762	0.256	0.077	0.034	0.039	**1.143**	0.070	29(3)
95	0.056		0.002	0.005			0.005	0.111	0.008	0.002	0.108	0.111	0.124	0.016	0.167	**0.462**	39(6)
PA (SE)	92 (2)	99 (1)	44(4)	70(4)	75(6)	81(7)	85(7)	73(2)	82(2)	49(5)	89(1)	86(2)	63(4)	90(2)	67(5)	51(7)	

for each of NLCD 2001 and NLCD 2006 (Tables 4 and 5). Additionally, developed open space accounted for approximately 17% of map disagreement for each of NLCD 2001 and NLCD 2006. The low density of impervious surface in this class (<20%) translates into a high proportion of grass. Either the urban context of the vegetation in this class was not evident during the photointerpretation of reference labels or the NLCD maps have misrepresented the urban context. Relatedly, there was an apparent increase in user's accuracies for the four developed classes (classes 21 through 24). User's accuracies for the developed classes increased as the level of urbanization (i.e., percentage impervious surface) increased, with the largest increase (~20%) occurring from class 23 to class 24. Regionally, overall accuracies were variable, ranging from 60% (region 9) to 90% (region 2) for NLCD 2001 and NLCD 2006, and, as expected, restricting the definition of agreement to a match between the map label and primary label only reduced overall agreement substantially relative to agreement defined as a match between the map label and the primary or alternate reference label (Supplementary material, Table S1).

Aggregating the NLCD classes to Level I improved overall accuracy from 79% to 85% for NLCD 2001 and from 78% to 84% for NLCD 2006 (Table 6). The modest improvement (~6%) in overall accuracy realized from class aggregation is an indication that much of the disagreement was not contained within classes in the same hierarchical group. Regional variation in overall accuracy at Level I (Supplementary material, Table S2) was not as dramatic as it was at Level II (Supplementary material, Table S1), and the increase in accuracy realized by inclusion of the alternate reference label was also less dramatic at Level I than Level II.

The procedures used for the NLCD 2001–2006 mapping resulted in small changes to the NLCD 2001, and these changes have been acknowledged (Fry et al., 2011, p. 859). Two versions of NLCD 2001 land-cover data now exist. The NLCD 2001 data in the NLCD 2001–2006 product are clearly labeled as version 2 (http://www.mrlc.gov/nlcd01_data.php). Thus, the accuracy results reported in Wickham et al. (2010) are an assessment of NLCD 2001 version 1, whereas the results reported here assess NLCD 2001 version 2. Users of the original (version 1) NLCD 2001 land-cover product (Homer et al., 2007) should continue to refer to the Wickham et al. (2010) accuracy results, whereas users of the newer NLCD 2001 product (Fry et al., 2011) should refer to the accuracy results reported in this article. The NLCD 2001 Level II and

Table 5
NLCD 2006 Level II error matrix. Cell entries are expressed as percent of area, and based on agreement defined as a match between the map class and either the primary or alternate reference class. User's accuracy (UA) and producer's accuracy (PA) are reported with standard errors (SE) in parentheses. Overall accuracy is 78% (0.8%) Main diagonal entries (agreement) are in bold.

Map	Reference																
	11	12	21	22	23	24	31	41	42	43	52	71	81	82	90	95	UA (SE)
11	**1.988**		0.016				0.007	0.007			0.021	0.029	0.001	0.021	0.024	0.003	94(2)
12		**0.006**					0.001					0.004					53(2)
21	0.017		**1.615**	0.194	0.029	0.009	0.014	0.272	0.126	0.105	0.138	0.142	0.224	0.203	0.029		52(2)
22	0.011		0.339	**0.916**	0.096	0.044	0.014	0.018	0.010		0.050	0.012	0.015	0.038	0.005		59(2)
23	0.021		0.107	0.052	**0.656**	0.074	0.003		0.005			0.022	0.009				69(2)
24				0.009		**0.045**											83(2)
31	0.034		0.014	0.003	0.001	0.003	**0.828**	0.010	0.029	0.002	0.143	0.135	0.009	0.006	0.001	0.006	68(2)
41			0.175	0.034				**9.799**	0.806	0.404	0.318	0.088	0.152	0.144	0.196		81(2)
42			0.125					0.733	**10.266**	0.278	0.729	0.136	0.003	0.098	0.001	0.001	83(2)
43			0.018					0.081	0.001	**1.616**	0.106	0.060			0.024		85(2)
52	0.013		0.446	0.022	0.006		0.096	0.599	0.534	0.283	**18.751**	0.546	0.390	0.146	0.012	0.154	85(2)
71	0.010		0.346	0.039			0.004	0.253	0.187	0.032	0.451	**11.052**	1.569	0.710	0.006	0.124	75(2)
81			0.292	0.032			0.017	0.202	0.032		0.411	0.383	**5.130**	0.083	0.007	0.048	77(2)
82	0.001		0.376	0.007	0.007			0.145	0.033	0.001	0.165	0.223	0.879	**14.080**	0.055	0.019	88(2)
90	0.026		0.020		0.030			1.150	0.341	0.716	0.325	0.080	0.038	0.045	**1.149**	0.056	29(2)
95	0.061		0.003	0.005				0.116	0.008	0.002	0.109	0.107	0.128	0.022	0.161	**0.467**	39(2)
PA (SE)	91(2)	100(0)	42(4)	70(4)	80(5)	26(10)	84(7)	73(2)	83(2)	47(5)	86(1)	85(2)	60(4)	90(2)	69(6)	53(11)	

Table 6
Level I error matrices for NLCD 2001 (top) and NLCD 2006 (bottom) for the continental United States. Cell entries are percentage of area and agreement is defined as a match between the map class and either the primary or alternate reference class. User's accuracy (UA) and producer's accuracy (PA) are reported with standard errors (SE) in parentheses. Overall accuracy is 85% (0.7%) for 2001 and 84% (0.7%) for 2006. Main diagonal entries (agreement) are in bold.

2001	Reference								
Map	10	20	30	40	50	70	80	90	UA (SE)
10	**2.005**	0.016	0.002	0.007	0.008	0.023	0.037	0.054	93 (1)
20	0.005	**4.039**	0.018	0.545	0.188	0.187	0.482	0.082	73 (2)
30	0.039	0.015	**0.831**	0.057	0.115	0.129	0.016	0.009	69 (4)
40	0.001	0.201		**25.414**	0.623	0.230	0.222	0.055	95 (1)
50	0.008	0.373	0.095	1.502	**18.517**	0.580	0.605	0.166	85 (2)
70	0.007	0.379	0.004	0.509	0.372	**11.055**	2.170	0.119	76 (3)
80	0.040	0.536	0.019	0.231	0.419	0.312	**21.122**	0.064	93 (1)
90	0.062	0.053	0.005	2.375	0.310	0.181	0.212	**1.939**	38 (3)
PA (SE)	93 (2)	72 (4)	85 (7)	83 (1)	90 (1)	87 (2)	85 (2)	78 (5)	
2006									
10	**1.994**	0.016	0.008	0.007	0.021	0.033	0.022	0.027	94 (1)
20	0.036	**4.232**	0.030	0.524	0.186	0.169	0.480	0.035	74 (2)
30	0.034	0.019	**0.828**	0.041	0.143	0.135	0.015	0.007	68 (4)
40		0.230		**24.671**	0.891	0.223	0.327	0.055	93 (1)
50	0.013	0.474	0.096	1.416	**18.755**	0.546	0.536	0.166	85 (2)
70	0.010	0.385	0.004	0.473	0.452	**11.032**	2.280	0.130	75 (3)
80	0.001	0.695	0.017	0.318	0.569	0.430	**20.527**	0.075	91 (1)
90	0.062	0.058	0.001	2.295	0.384	0.174	0.232	**1.953**	38 (3)
PA (SE)	93 (2)	74 (4)	68 (7)	94 (1)	85 (1)	75 (2)	91 (2)	39 (6)	

Level I overall accuracies reported here (version 2) are generally consistent with the results reported by Wickham et al. (2010).

Overall accuracy of the binary change/no change classification exceeded 95% across all regions (Table 7) when using the primary and alternate reference changes labels (attributes 15 through 18 in Table 3) to define agreement. The high overall agreement rate was driven by the large proportion of area of no change. User's and producer's accuracies for no-change were consistently above 95%, whereas user's and producer's accuracies for change were lower and more variable over the 10 regions. User's accuracies for the change varied regionally from 69% to 88%, whereas producer's accuracies for change were generally much lower, below 50% for some regions. Constraining the definition of agreement to a match between the map label and the primary reference label only (attribute 15 in Table 3) did not reduce overall agreement substantially, but had a noticeable impact on user's and producer's accuracies for the change class (Supplementary material, Table S3). On average, constraining the definition of agreement to a match between the map label and the primary reference label only reduced user's and producer's accuracies for the change class by ~26% and ~37%, respectively, whereas user's and producer's accuracies for the no change class were reduced by only ~1% and ~3%, respectively.

Nationally, user's accuracies for the reporting themes ranged from 27% to 93% (Table 8). Overall, the no-change reporting themes had higher user's accuracies than the change reporting themes. User's

Table 7
Error matrix and accuracy statistics (expressed as percents) for a binary change and no-change classification where agreement is defined as a match between the sampled pixel map label and any of the four reference change labels (i.e., attributes 15 through 18 in Table 3). Overall accuracy is in bold at the intersection of the Producer's Acc row and User's Acc column. Values in parentheses are standard errors.

Map	Reference change	No change	User's Acc	Map	Reference change	No change	User's Acc
Region 1				Region 6			
Change	2.320	0.421	85 (2)	Change	0.552	0.157	78 (2)
No change	0.300	96.958	100 (0)	No change	0.781	98.509	99 (0)
Producer's Acc	89 (5)	100 (0)	**99 (0)**	Producer's Acc	41 (14)	100 (0)	**99 (0)**
Region 2				Region 7			
Change	0.886	0.134	87 (1)	Change	3.074	0.423	88 (2)
No change	0.309	98.671	100 (0)	No change	1.950	94.553	98 (1)
Producer's Acc	74 (11)	100 (0)	**99 (0)**	Producer's Acc	61 (10)	100 (0)	**98 (1)**
Region 3				Region 8			
Change	0.911	0.135	87 (1)	Change	2.399	0.376	86 (3)
No change	1.584	97.370	98 (1)	No change	0.732	96.493	99 (0)
Producer's Acc	37 (13)	100 (0)	**98 (1)**	Producer's Acc	77 (7)	100 (0)	**99 (0)**
Region 4				Region 9			
Change	0.617	0.260	70 (2)	Change	5.163	0.676	88 (1)
No change	0.428	98.695	100 (0)	No change	3.536	90.626	96 (1)
Producer's Acc	59 (16)	100 (0)	**99 (0)**	Producer's Acc	59 (7)	99 (1)	**96 (1)**
Region 5				Region 10			
Change	0.393	0.107	79 (2)	Change	0.759	0.336	69 (3)
No change	1.491	98.008	99 (1)	No change	0.012	98.892	100 (0)
Producer's Acc	21 (9)	100 (0)	**98 (1)**	Producer's Acc	98 (0)	100 (0)	**100 (0)**

National Map	Reference change	No change	User's Acc
Change	1.434	0.263	84.5 (0.6)
No change	1.064	97.239	98.9 (0.2)
Producer's Acc	57.4 (4.6)	99.7 (0.0)	98.7 (0.2)

Table 8
Regional user's accuracies (%) for reporting themes identified in Table 2b (standard error for National estimate in parentheses).

	Regions										
Class	1	2	3	4	5	6	7	8	9	10	National
Water loss	40	92	80	70	84	78	80	64	68	58	80 (2)
Water gain	80	52	82	76	78	52	82	88	86	76	76 (2)
Urban gain	73	80	81	73	75	72	67	69	73	58	72 (1)
Forest loss	83	76	88	72	60	66	82	80	89	62	82 (1)
Forest gain	47	16	32	29	35	32	71	79	75	37	69 (2)
Shrub loss	55	77	49	45	4	37	70	69	79	37	66 (3)
Shrub gain	68	52	76	59	47	67	67	80	66	47	68 (2)
Grassland loss	72	71	44	48	37	46	63	56	58	62	57 (2)
Grassland gain	88	88	73	57	52	69	67	53	75	58	69 (2)
Agriculture loss	41	35	30	26	58	54	43	32	35	20	39 (1)
Agriculture gain	33	40	27	37	27	16	34	24	9	12	27 (2)
Water, no change	100	96	96	64	88	88	100	80	100	100	93 (2)
Urban, no change	69	75	67	71	73	75	69	77	72	73	73 (2)
Forest, no change	96	76	92	88	91	91	95	99	93	96	93 (1)
Shrub., no change	69	95	89	80	23	27	24	16	20	31	85 (2)
Grass., no change	94	97	96	71	72	48	60	18	36	31	75 (3)
Agric., no change	89	88	84	96	96	92	84	72	87	65	91 (1)

accuracies for the no-change reporting themes ranged from 73% (urban, no change) to 93% (water, no change; forest, no change). User's accuracies were high for the reporting themes water loss (80%), water gain (76%), urban gain (72%), and forest loss (82%) and only three of the change reporting themes had user's accuracies below 65%—agriculture gain (27%), agriculture loss (39%), and grassland loss (57%). As expected, user's accuracies varied by region. For example, the forest gain reporting theme had much higher user's accuracies in the eastern regions 7, 8, and 9 than in the other seven regions (1–6, 10). Producer's accuracies were generally lower than user's accuracies. For the no change reporting themes, producer's accuracies ranged from 59% (no change, grassland) to 86% (no change, water), and never exceeded 40% for the change reporting themes (Table 9). User's accuracy for increase in impervious surface was 67% (Table 10), which was consistent with the user's accuracy for the urban gain reporting theme. However, the producer's accuracy for increase in impervious surface was very low (8%).

Additional analyses examined relationships between accuracy and map heterogeneity, accuracy and reference label confidence, and accuracy and the difference between the map and reference data acquisition dates. Map heterogeneity had the expected effect of reducing agreement (Supplementary material, Table S4). The highest agreement always occurred when there was only one map land-cover class within a 3-by-3 pixel window surrounding the sample pixel. Similarly, accuracy declined as the confidence in the reference label assignment declined (Supplementary material, Table S5). Differences between map and reference acquisition dates did not have a significant effect on agreement results. Significant differences (two-sample t-test, $\alpha = 0.05$) were found in only 9 of 30 comparisons (Supplementary material, Table S6) and for 8 of the 9 significant differences there was a greater difference between map and reference acquisition dates for map and reference labels that matched than map and reference labels did not match. These results are consistent with map versus reference acquisition date accuracy comparisons reported in Wickham et al. (2010). The average difference in Landsat and reference acquisition dates was ~3 years for 2001 and ~7 months for 2006, reflecting the wider availability of high resolution digital reference imagery for 2006.

4. Discussion

National overall accuracies for NLCD 2001 and NLCD 2006 approached 80%, and user's accuracies for all upland forests classes (41–43), shrubland (52), and cropland (82) exceeded 80%. These single date accuracies translated into high overall accuracies across all regions for the binary change–no change classification (Table 7), reasonably high user's accuracies for the change reporting themes water gain, water loss, forest loss and urban gain, and high user's accuracies for the no-change reporting themes for water, urban, forest, and agriculture.

Table 9
Regional producer's accuracies (%) for reporting themes identified in Table 2b (standard error for National estimate in parentheses).

	Regions										
Class	1	2	3	4	5	6	7	8	9	10	National
Water loss	16	76	73	58	32	71	18	11	30	69	39 (10)
Water gain	57	55	21	51	13	5	14	74	77	8	21 (4)
Urban gain	39	48	3	86	6	30	16	24	21	38	18 (4)
Forest loss	44	22	32	77	4	13	31	36	29	61	30 (3)
Forest gain	36	2	88	79	4	1	35	47	41	26	32 (4)
Shrub loss	28	65	8	55	0	1	33	11	28	19	23 (3)
Shrub gain	36	23	20	22	1	4	13	18	14	19	18 (2)
Grassland loss	34	10	8	43	7	11	13	27	22	7	18 (2)
Grassland gain	45	48	9	70	4	64	46	39	47	39	34 (6)
Agriculture loss	22	17	5	8	12	23	9	13	16	19	11 (2)
Agriculture gain	92	27	31	29	4	100	4	11	13	68	12 (4)
Water, no change	91	93	96	77	88	100	85	85	73	76	86 (3)
Urban, no change	77	28	42	54	74	69	67	73	72	58	63 (4)
Forest, no change	80	87	93	49	91	66	76	90	57	85	79 (1)
Shrub., no change	68	92	81	58	3	8	31	13	34	25	74 (2)
Grass., no change	74	18	46	84	95	23	52	21	25	2	59 (2)
Agric., no change	74	89	85	55	88	92	87	99	83	91	79 (2)

Table 10
Accuracy of change in percentage impervious surface for the continental United States. Increase is denoted with the symbol ↑, and no change is denoted with the symbol "no Δ". User's and producer's accuracy columns are abbreviated UA and PA respectively. Main diagonal entries are in bold, overall agreement is reported in bold italics at the intersection of the UA column and PA row, and standard errors are in parentheses.

	Reference		
Map	%IS ↑	No Δ %IS	UA
%IS ↑	**0.1009**	0.0502	66.8 (1.2)
No Δ %IS	1.1461	**98.7028**	98.9 (0.2)
PA	8.1 (1.2)	99.9 (0.0)	**98.8 (0.1)**

Producer's accuracies for the change classes were lower than the corresponding user's accuracies (compare Table 8 and Table 9).

The motivation for NLCD mapping and product development is to supply data needed by the user community. Accuracy assessment is a standard component of the NLCD mapping protocol because the user community needs information on data quality to guide their work. For example, cropland user's and producer's accuracies are 88% to 90% for both 2001 and 2006 NLCD. Distinguishing cropland from pasture is important to many water-quality studies because fertilizer and pesticides are predominantly applied to croplands rather than to pastures, and therefore confusion between cropland and pasture would limit the use of NLCD data for assigning fertilizer and pesticide applications across the landscape (Nakagaki & Wolock, 2005). The single-date accuracy assessment results indicate that the data are of sufficient quality to support many user-community applications. The change accuracy assessment results also indicate that data quality for most of the no-change classes, water loss and gain, forest loss and urban gain will support many uses. Deforestation, urbanization, and flux in the amount of water in the environment are recognized as important processes that directly relate to environmental quality (S.V. Smith, et al., 2002; MEA, 2005; McKinney, 2006; http://www.epa.gov/roe).

Common sources of disagreement were attributable to determining the context of grass and distinguishing class 21 (developed, open space) from most other classes. These classification errors propagated to misclassification within the associated change classes. For example, despite high user's accuracies for cropland and high user's accuracies for no change agriculture, user's accuracies for the agriculture gain and agriculture loss reporting themes were very low. Intuitively, this inconsistency is partly attributable to the general difficulty in assigning the context of grass for the individual eras because accurate detection of change relies on accurate classification at both dates.

The NLCD 2001 and NLCD 2006 single date class-specific accuracies were not universally reliable predictors of class-specific accuracies for the 2001 to 2006 change and no change classes. For example, 2001 and 2006 Level I user's and producer's accuracies for agriculture ranged from 84% to 93% (Table 6) and user's and producer's accuracies for the no change agriculture class were 91% and 79% (Tables 8 and 9) respectively, but user's and producer's accuracies for the agriculture gain and loss classes were less than 40%. van Oort (2007) provided statistical methods for estimating the accuracy of change error matrices from single-date error matrices under assumptions of independence or positive correlation of classification errors for the two dates. Our results suggest that relationships between single-date and change accuracies are complex and class-specific rather than map-specific.

Of the three main components of accuracy assessment (sample design, response design, analysis) (Stehman & Czaplewski, 1998), change detection accuracy assessment research has been focused primarily on the sampling design (e.g., Biging et al., 1999; Stehman, 2012) and analysis components (Burnicki, 2011; Burnicki et al., 2010; Foody, 2009; Hester et al., 2010; Li & Zhou, 2009; Liu & Zhou, 2004; van Oort, 2007). There has been little research on how response design protocols (the procedures for obtaining the reference classification) should be adjusted as the main objective changes from land-cover accuracy assessment to land-cover change accuracy assessment. Our results indicate that additional research on response design protocols for land-cover change accuracy assessments would complement the important contributions from the emerging research on change detection accuracy assessment. Some potential research topics are:

- *Evaluate different response design protocols for determining reference land cover change*. Our main response design protocols included primary and alternate labels for individual dates which were used to derive four reference land-cover change labels; many other response design protocols could be used, and studies comparing results from two or more response design protocols would likely advance this aspect of land-cover change accuracy assessment.
- *Evaluate photointerpeters' ability to detect land-cover change from high resolution reference media*. To our knowledge there are no formalized studies evaluating reference data quality in the context of land-cover change. Our results indicate that photointerpreters were often uncertain about whether land-cover change had actually occurred (Table S5), suggesting that more research is needed on identifying land-cover change from high resolution digital reference media. It does not appear that evaluation of reference data quality has been incorporated in many previous land-cover change accuracy assessment studies (e.g., Burnicki, 2011; Hester et al., 2010; Li & Zhou, 2009; Liu & Zhou, 2004)
- *Extend analysis of sensitivity of reference label assignment to the interpreter assigning the label for single-date maps to the temporal domain*. Consistency of reference label assignment is affected by class rarity and class fuzziness (Lunetta et al., 2001; Mann & Rothley, 2006; Powell et al., 2004). It is likely that the geometric growth in the number of classes that occurs as one moves from land-cover mapping to land-cover change mapping compounds these problems. Quantification of the impact of class rarity and class fuzziness on land-cover change reference label assignment would help to document the degree of difficulty in obtaining confident and consistent reference land-cover change labels.

The multi-objective nature of the NLCD 2006 assessment also raises research questions related to sampling design for simultaneously assessing accuracy of land-cover change and accuracy of the maps for the dates bracketing the change period. The sampling design implemented in this assessment was stratified to increase the sample size for the rarer land-cover change classes. An impact of this choice of stratification is that standard errors for the accuracy estimates are higher for the single date (2001 and 2006) assessments than would have been the case had the stratification been based on the map land-cover classes of either date. The questions of how to choose strata and how to allocate the sample sizes to the chosen strata are an important sampling design consideration when the assessment must satisfy the multiple objectives of accuracy assessment of both dates and accuracy assessment of change.

The stratified estimators require weighting sample pixels by the inverse of their inclusion probabilities. These estimation weights can vary enormously among strata. Sample pixels from the no change strata had high estimation weights because these strata had low inclusion probabilities. For example, in Region 1 approximately 222 million pixels were mapped as no change forest and 75 were sampled yielding an estimation weight of nearly 3 million for each pixel sampled from this stratum. In contrast, a sample pixel from the forest to urban change stratum in Region 1 had an estimation weight of approximately 1000. Because of the necessary weighting feature of the estimators used with the stratified design, a change omission error for a sampled pixel from a stratum with high estimation weights (e.g., the no change forest stratum in Region 1) would have an enormous influence on estimated producer's accuracy. Stratification is essentially a prerequisite for land-cover change accuracy assessment because the relatively rare change classes will have small sample sizes without stratification. Our results suggest that there may be a stronger coupling between sampling and response design components for land-cover change accuracy assessment than land-cover accuracy assessment, such that obtaining accurate reference label assignments for sample pixels in strata with high estimation weights (e.g., the no change strata in Table 2) is essential. More research is needed on linkages between sampling design and response design components for land-cover change accuracy assessment.

In addition to documentation of data quality to help inform the user community, another important aspect of accuracy assessment is to inform and guide future map production. The results of the NLCD 2006 change detection accuracy assessment suggest that mapping protocols need to be further developed and refined to better distinguish the context of grass. Disagreement among grass-dominated classes accounted for approximately 26% of the error in the single-date classifications, and this level of error constrains the level of change accuracy that can be realized. The MRLC consortium continues to develop its integrated (among federal agencies) model for NLCD data production, and the continued alignment of the USDA-NASS Cropland Data Layer (CDL) (Johnson & Mueller, 2010) with NLCD products will in all likelihood improve distinction between grassland, pasture, and cropland. Such continued integration should improve change accuracies among these classes in future NLCD products. An additional methodological improvement would be to further integrate NLCD accuracy assessments with NLCD map production. Use of periodic accuracy assessments on a limited geographic scale during the mapping process would likely lead to improvement in identification of land-cover change from reference imagery and this improvement would likely feedback to the mapping process. Such a protocol might be difficult to incorporate into NLCD mapping because of the short five-year interval between mapping eras and the continental extent of the mapped area; however, investigation of its feasibility appears to be warranted.

5. Summary

NLCD 2006 provides the first wall-to-wall land-cover change database for the continental US, and the accuracy assessment of NLCD 2006 reported herein provides the first known comprehensive evaluation of a continental, Landsat-based land-cover change database. The following points summarize our main findings.

- NLCD provides high quality data for static assessments that require land-cover data and dynamic assessments related to deforestation (e.g., Riitters & Wickham, 2012), urbanization, and flux in the amount of water.
- High overall and user's accuracies for the individual dates translated into high user's accuracies for the 2001–2006 change reporting themes water gain and loss, forest loss, urban gain, and the no-change reporting themes for water, urban, forest, and agriculture.
- The main factor limiting higher accuracies for the change reporting themes appeared to be difficulty in distinguishing the context of grass. Confusion among grass-dominated categories accounted for 26% of the classification errors in NLCD 2001 and in NLCD 2006.
- Improvements in NLCD single date class-specific accuracies are needed to improve land-cover class change accuracies in future NLCD products. Refinements to the NLCD mapping process (Fry et al., 2011 p. 862) and continued alignment of NLCD and USDA-NASS CDL data should improve single-date class-specific accuracies for some grass-dominated classes.
- Land-cover change accuracy assessment methodology is at a nascent stage of development relative to land-cover change mapping methodology.
- Response design protocols for land-cover change need further development. Research on the confidence and reliability of identification of land-cover change from photointerpretation of reference media and the impact of class rarity and fuzziness on the consistency of reference label assignment are lacking.
- Interaction between sampling design and response design may be more tightly connected for land-cover change accuracy assessment than land-cover accuracy assessment. Stratification is strongly motivated for land-cover change accuracy assessment because most land-cover change classes will be rare. A consequence of stratified sampling with sample size allocated disproportionately to strata of rare classes is that very accurate reference land-cover labels are required for sample pixels from the more common classes (i.e., high estimation weight strata).

Acknowledgments

The United States Geological Survey (USGS) provided primary funding for this research. The United States Environmental Protection Agency through its Office of Research and Development partially funded and collaborated in the research described here. It has been subjected to Agency review and approved for publication. Stehman was supported by Grant/Cooperative Agreement Number G12AC20221 from the USGS. Lori Baer, Jesse Ballew, Brett Bunde, Adam Case, Catherine Costello, Michelle Coffey, Allison Paraham, Rick Poss, Dan Sorenson, and Gail Winters of USGS collected the reference data. We thank two anonymous reviewers for their constructive suggestions to improve the paper. Any use of trade, form, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.rse.2012.12.001.

References

Biging, G. S., Colby, D. R., & Congalton, R. G. (1998). 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.

Burnicki, A. C. (2011). Modeling the probability of misclassification in a map of land cover change. *Photogrammetric Engineering and Remote Sensing*, *77*, 39–49.

Burnicki, A. C., Brown, D. G., & Goovaerts, P. (2010). Propagating error in land-cover change analyses: Impact of temporal dependence under increased thematic complexity. *International Journal of Geographical Information Science*, *24*, 1043–1060.

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

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., Xian, G., Jin, S., Dewitz, J. A., Homer, C. G., Yang, L., et al. (2011). Completion of the 2006 National Land Cover Database for the conterminous United States. *Photogrammetric Engineering and Remote Sensing*, *77*, 858–864.

Gallego, F. J. (2012). The efficiency of sampling very high resolution images for area estimation in the European Union. *International Journal of Remote Sensing*, *33*, 1868–1880.

Gopal, S., & Woodcock, C. (1994). Theory and methods for accuracy assessment of thematic maps using fuzzy sets. *Photogrammetric Engineering and Remote Sensing*, *60*, 181–188.

Hester, D. B., Nelson, S. A. C., Cakir, H. I., Khorram, S., & Cheshire, H. (2010). High-resolution land cover change detection based on fuzzy uncertainty analysis and change reasoning. *International Journal of Remote Sensing*, *31*, 455–475.

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.

Johnson, D. M., & Mueller, R. (2010). The 2009 cropland data layer. *Photogrammetric Engineering and Remote Sensing*, *76*, 1201–1205.

Li, B., & Zhou, Q. (2009). Accuracy assessment on multi-temporal land-cover change detection using a trajectory error matrix. *International Journal of Remote Sensing*, *30*, 1283–1296.

Liu, H., & Zhou, Q. (2004). Accuracy analysis of remote sensing change detection by rule-based rationality evaluation with post classification comparison. *International Journal of Remote Sensing*, *25*, 1037–1050.

Lunetta, R. S., Iames, 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 set points. *International Journal of Remote Sensing*, *27*, 5027–5036.

McKinney, M. L. (2006). Urbanization as a major cause of biotic homogenization. *Biological Conservation*, *127*, 247–260.

Millennium Ecosystem Assessment (MEA) (2005). *Ecosystems and human well-being: Synthesis.* Washington, D.C.: Island Press.

Nakagaki, N., & Wolock, D. M. (2005). *Estimation of agricultural pesticide use using land cover maps and county pesticide data.* U.S. Geological Survey Open-File Report 2005-1188 (available at: http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA438834).

Olofsson, P., Stehman, S. V., Woodcock, C. E., Sulla-Menashe, D., Sibley, A. M., Newell, J. D., et al. (2012). A global land-cover validation data set, part I: Fundamental design principles. *International Journal of Remote Sensing*, *33*, 5768–5788.

Pontius, R. G., Jr., & Cheuk, M. L. (2006). A generalized cross-tabulation matrix to compare soft-classified maps at multiple spatial resolutions. *International Journal of Geographical Information Science*, *20*, 1–30.

Powell, R. L., Matzke, N., de Souza, 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.

Riitters, K. H., & Wickham, J. D. (2012). Decline of forest interior conditions in the conterminous United States. *Scientific Reports*. http://dx.doi.org/10.1038/srep00653.

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

Smith, S. V., Renwick, W. H., Bartley, J. D., & Buddemeier, R. W. (2002). Distribution and significance of small artificial water bodies across the United States landscape. *The Science of the Total Environment*, *299*, 21–36.

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.

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.

Stehman, S. V. (1997). Estimating standard errors of accuracy assessment statistics under cluster sampling. *Remote Sensing of Environment*, *60*, 258–269.

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., Olofsson, P., Woodcock, C. E., Herold, M., & Friedl, M. A. (2012). A global land cover validation dataset, II: Augmenting a stratified sampling design to estimate accuracy by region and land-cover class. *International Journal of Remote Sensing*, *33*, 6975–6993.

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. (2011). Pixels, blocks of pixels, and polygons: Choosing a spatial unit for thematic accuracy assessment. *Remote Sensing of Environment*, *115*, 3044–3055.

Stehman, S. V., Wickham, J. D., Fattorini, L., Wade, T. G., Baffetta, F., & Smith, J. H. (2009). Estimating accuracy of land-cover composition from a two-stage cluster design. *Remote Sensing of Environment*, *113*, 1236–1249.

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.

van Oort, P. (2007). Interpreting the change detection error matrix. *Remote Sensing of Environment*, *108*, 1–8.

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.

Wickham, J. D., Stehman, S. V., Fry, J. A., Smith, J. H., & Homer, C. G. (2010). Thematic accuracy of the NLCD 2001 land cover for the conterminous United States. *Remote Sensing of Environment*, *114*, 1286–1296.

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.

Xian, G., Homer, C., Dewitz, J., Fry, J., Hossain, N., & Wickham, J. (2011). Change of impervious surface area between 2001 and 2006 in the conterminous United States. *Photogrammetric Engineering and Remote Sensing*, *77*, 758–762.

Xian, G., Homer, C., & Fry, J. (2009). Updating the 2001 National Land Cover Database land cover classification to 2006 by using Landsat imagery change detection methods. *Remote Sensing of Environment*, *113*, 1133–1147.