Optimizing selection of training and auxiliary data for operational land cover classification for the LCMAP initiative

Zhe Zhu ^{a,e,*}, Alisa L. Gallant ^b, Curtis E. Woodcock ^c, Bruce Pengra ^d, Pontus Olofsson ^c, Thomas R. Loveland ^b, Suming Jin ^c, Devendra Dahal ^d, Limin Yang ^d, Roger F. Auch ^b

^a Department of Geosciences, MS 1053, Science Building 125, Texas Tech University, Lubbock, TX 79409, USA

^b USGS, EROS Center, 47914 252nd Street, Sioux Falls, SD 57198, USA

^c Department of Earth and Environment, Boston University, 685 Commonwealth Avenue, Boston, MA 02215, USA

^d Stinger Ghaffarian Technologies (SGT), Contractor to the USGS, EROS Center, 47914 252nd Street, Sioux Falls, SD 57198, USA

^e ASRC InuTeq, Contractor to the U.S. Geological Survey (USGS), Earth Resources Observation and Science (EROS) Center, 47914 252nd Street, Sioux Falls, SD 57198, USA

* Corresponding author: Zhe Zhu; E-mail address: zhe.zhu@ttu.edu

ABSTRACT:

The U.S. Geological Survey's Land Change Monitoring, Assessment, and Projection (LCMAP) initiative is a new end-to-end capability to continuously track and characterize changes in land cover, use, and condition to better support research and applications relevant to resource management and environmental change. Among the LCMAP product suite, there are annual land cover maps that will be available to the public. This paper describes an approach to optimize the selection of training and auxiliary data for deriving the thematic land cover maps based on all available clear observations from Landsats 4–8. Training data were selected from map products of the U.S. Geological Survey's Land Cover Trends

project. The Random Forest classifier was applied for different classification scenarios based on the Continuous Change Detection and Classification (CCDC) algorithm. We found that extracting training data proportionally to the occurrence of land cover classes was superior to an equal distribution of training data per class, and recommend a total of 20,000 training pixels to classify an area about the size of a Landsat scene. The problem of unbalanced training data was alleviated by extracting a minimum of 600 training pixels and a maximum of 8,000 training pixels per class. We additionally explored removing outliers contained within the training data based on their spectral and spatial criteria, but observed no significant improvement in classification results. We also tested the importance of different types of auxiliary data that were available for the conterminous United States, including: (a) five variables used by the National Land Cover Database, (b) three variables from the cloud screening “Function of mask” (Fmask) statistics, and (c) two variables from the change detection results of CCDC. We found that auxiliary variables such as a Digital Elevation Model and its derivatives (aspect, position index, and slope), potential wetland index, water probability, snow probability, and cloud probability improved the accuracy of land cover classification. Compared to the original strategy of the CCDC algorithm (500 pixels per class), the use of the optimal strategy improved the classification accuracies substantially (15 percentage points increase in overall accuracy and 4 percentage points increase in minimum accuracy).

Keywords: Continuous Change Detection and Classification (CCDC), training strategy, auxiliary data, land cover classification, Landsat

1. Introduction

The U.S. Geological Survey (USGS) has a long history in land cover investigations starting with the landmark paper, “A Land Use and Land Cover Classification System for use with Remote Sensor Data” (Anderson et al., 1976), and including global land cover mapping (Loveland et al., 2000) and the ongoing production of the National Land Cover Database (NLCD; Vogelmann et al., 2001; Homer et al., 2007, 2015; Fry et al., 2011). Whereas these past projects have had a significant impact on the data user

community, data needs are changing and innovative products and mapping approaches to characterize land cover are required to meet the community's appetite for science-quality, geospatial land cover data.

The emergence of land change science has had a particularly significant influence on land cover mapping and monitoring requirements. Driven by concerns about climate change and resource sustainability, Rindfuss et al. (2004) and Turner et al. (2007) defined land change science as a framework to understand the interactions between people and nature that lead to changes in the type, magnitude, condition, and location of land use and cover. Implicit is the need to use remote sensing, modeling, and synthesis to understand and explain land change rates, causes, and consequences. Meeting the diverse needs of land cover data users means that the data producers must transform traditional approaches into new strategies that generate higher quality products with additional land cover variables, more detailed legends, and more frequent geospatial information and statistics on change.

The USGS response to the growing requirements for land use, cover, and condition data, information, and knowledge is the Land Change Monitoring, Assessment, and Projection (LCMAP) initiative. LCMAP is an end-to-end capability to continuously track and characterize changes in land cover, use, and condition and translate such information into assessments of current and historical processes of cover and change that can serve as the science foundation for research and applications relevant to resource management and environmental change. As implied in the name, key elements include mapping and monitoring change, analyzing and assessing the rates, causes, and consequences of change, and providing scenario-based projections of future land change. LCMAP continues the long-standing USGS commitment to land cover research and development and represents the foundation for a Landsat-based Federal land monitoring system. Furthermore, LCMAP will stimulate geographic research on understanding the connections between human activity and natural systems, and ultimately represents an essential step towards better understanding the combined impacts of climate and land use and land cover change.

The current study addresses the “M” in LCMAP, which aims to generate science-quality land cover and land change products from current and near-real time Earth observations. Landsat data are the primary input, due to their relatively dense temporal coverage (e.g., 8 or 16 days) and global coverage going back to 1972 (Wulder et al., 2015). LCMAP is initially focused on the Landsat 4–8 era over the United States, but MSS data from Landsats 1–3 will eventually be included, and the ultimate geographic scope of LCMAP will be expanded to the globe.

LCMAP uses all available Landsat data for any given location to characterize historical land change at any point across the full Landsat record and to detect and characterize land cover and land change as it occurs. The basis for this is the Continuous Change Detection and Classification (CCDC) algorithm (Zhu & Woodcock, 2014b; Zhu et al., 2015b). The initial plan for LCMAP land cover and land change products are annual land cover maps and maps of key change attributes (e.g., date, type, and magnitude of change). This article focuses on a strategy for the operational, automated generation of annual land cover maps, specifically on an optimized approach to train the classifier that produces the maps. The research presented here contributes to the LCMAP goal to produce annual land cover maps from 1985–2015, and eventually to the goal to produce land cover maps at any point in the Landsat record. Thus, the lessons learned and reported here represent a key component of a transformative approach for land cover mapping and monitoring.

Land cover classification is recognized as an important scientific goal by the remote sensing community because of the impacts of land cover on major environmental process and functions such as the energy balance, carbon dynamics, and the hydrologic cycle (Anderson et al., 1976; Loveland et al., 2000; Friedl et al., 2002; Wulder et al., 2008; Gong et al., 2013; Chen et al., 2015). The quest to provide accurate classification results has driven the development and testing of many classifiers, including maximum likelihood (Strahler, 1980), decision trees (Friedl et al., 2010), neural networks (Gopal et al., 1999), and support vector machines (Huang et al., 2002). Recently, the Random Forest Classifier (RFC; Breiman, 2001), a decisions tree approach, has been gaining wide use for remote sensing image classification (Ham

et al., 2005; Pal, 2005; Gislason et al., 2006; Lawrence et al., 2006; Cutler et al., 2007; Rodriguez-Galiano et al., 2012b; Zhu et al., 2012; Yan & Roy, 2015) for a number of reasons, including its ability to handle high dimensional and non-normally distributed data, its avoidance of overfitting, and the provision of measurements of variable importance (Breiman, 2001).

Advancements in classifiers have improved the accuracy of land cover maps generated from remotely sensed imagery (Huang et al., 2002; Lu & Weng, 2007; Gong et al., 2013). However, to further increase classification accuracy, more attention should be focused on the input data, as land cover classes are often not separable based only on multispectral data. Multitemporal and spatial information derived from remotely sensed data have been reported to be important in increasing land cover classification accuracy (Stuckens et al., 2000; Rodriguez-Galiano et al., 2012a; Zhu et al., 2012). Although better use of satellite data can result in higher classification accuracy, satellite data alone may not be sufficient to classify certain land cover features. For example, it is difficult to map wetlands solely based on satellite data, as the wetland designation is more a soil moisture regime than a cover type (Gallant, 2015). To help distinguish land cover features that are difficult to classify from satellite images, the use of auxiliary data has been recommended (Foody, 1995; Maselli et al., 1995; Vogelmann et al., 1998; Rogan et al., 2003; Wright & Gallant, 2007; Ozdogan & Gutman, 2008; Corcoran et al., 2013). There are many different kinds of auxiliary data that can be used to improve land cover classification, but their spatial and temporal limitations may hinder their usage.

Whereas the use of advanced classifiers and more informative input data can substantially improve classification accuracy, another factor – training of the classifier – is a critical step in supervised classification (Millard & Richardson, 2015). Foody and Arora (1997) found that the choice of training data had more impact on classification results than the choice of classifier. Moreover, in the current “data-boom” era, large amounts of potential training data are becoming available, and determining how to extract useful and representative training data merits attention.

The literature addressing strategies for selecting training data mainly centers on four aspects: distribution, size, balance, and outlier removal. Distribution refers to how the training data are distributed among the classes; the most frequently used distributions are *proportional*, where the amount of training data for a specific class is proportional to the size of that class, and *equal*, where the amount of training data is the same for all classes (Jin et al., 2014). Multiple studies have suggested that a proportional approach can achieve higher overall accuracy (Huang et al., 2002; Jin et al., 2014), but the proportions of the map classes in the study area need to be known prior to making the map. The amount of training data will also impact classification accuracy (Foody & Mathur, 2006), with the optimum size usually being related to the choice of classifier and dimensionality of the input data. For instance, the suggested amount of training data for the maximum likelihood classifier is at least 10–30 times the number of image features used (Piper, 1992; Mather, 2004). Generally, classification accuracy will increase with larger training data (Huang et al., 2002; Jin et al., 2014; Millard & Richardson, 2015). Sample balance is also important in training data selection, as unbalanced training data (i.e., dramatically more or less training data for one or multiple classes) may result in rare land cover types being under-represented relative to more abundant classes, which may degrade the overall classification accuracy (Weiss & Provost, 2003; Estabrooks et al., 2004; Mellor et al., 2015). Techniques such as down-sampling of majority classes (Freeman et al., 2012), over-sampling of minority classes (Ling and Li, 1998), and a combination of over-sampling and down-sampling training classes (Chawla et al., 2002) have been explored to alleviate the problem of unbalanced training data. Outliers in the training data may also impact the training process (Radoux et al., 2014) and can be eliminated based on spectral or spatial distances. For example, pixels that are spectrally far from the mean of the majority of pixels for a given class (such as measured by Mahalanobis distance) (Rousseeuw & Van Zomeren, 1990; Hodge & Austin, 2004) and pixels located close to the boundaries of different land cover classes can be identified as outliers (Radoux et al., 2014). Though a few studies have addressed different aspects of training data selection, we are unaware of a study that has fully explored all four aspects at the same time.

The CCDC algorithm, which is used to address the monitoring goals of LCMAP, applies a supervised Random Forest classifier for land cover classification. The classifier was trained using existing land cover maps generated by the USGS Land Cover Trends (LCT) project, which developed a national set of thematic maps following a methodology designed to achieve high quality, consistent results to support assessments of land change (Loveland et al., 2002). The LCT maps provide a locally precise and consistent source of national data for training operational land cover classification. However, it is unknown how to best select training data and appropriate auxiliary data to achieve the objectives of the LCMAP initiative. We conducted the current study to develop a robust strategy that answers the following questions:

- 1) What is the best way to distribute the training data among classes?
- 2) What is the optimum amount of training data?
- 3) What is the best balance of training data?
- 4) Is removing outliers in training data beneficial?
- 5) How much do different kinds of auxiliary data influence classification accuracy?

We address these questions in the order presented above, as the results build successively on the findings from one analysis to the next.

2. Study area

LCT data located in five Landsat scenes, which represented the ecoregional characteristics described in Table 1, were used to evaluate the different classification scenarios. These scenes were selected from a larger set of 14 path/rows that had previously been prepared for LCMAP product validation. The five scenes we chose offered the greatest variety of characteristics in land cover types, conditions, and drivers

of change in diverse environmental settings that also coincided with areas rich in data from the LCT project (Figure 1).

Table 1. Ecoregional characteristics of the five study areas.

Study Area (and path/row)	Ecoregion Characteristics
Puget Lowland and Cascades (46/27)	The west half of this landscape has fine-grained, interwoven, and highly dynamic landscape patterns resulting from an interspersed of low- to high-density urban development, agriculture, and forestry activities (Sleeter et al., 2012). The climate is cool and humid, with green biomass present year-round. The east half of the scene captures a portion of the Cascade Mountains and evidence of the long-active forest industry.
Central California (43/34)	Multi-cropping, perennial crops, and irrigation are among the fine-grained, complex agricultural practices blanketing the valley floor of this landscape punctuated by sprawling population centers (Sleeter et al., 2012). Surrounding foothills include some perennial crops, along with grazed annual grasslands and oak woodlands. The Mediterranean climate imposes a multimodal growing season in a water-limited landscape.
Flint Hills (28/33)	This landscape was originally tallgrass prairie, but now has land use that fluctuates between grassland and agriculture or has been developed. The region is naturally subject to high fire frequency followed by relatively short-lived fire scars, but fire suppression practices have allowed encroachment of trees into grasslands in some areas (Taylor et al., 2015).
Northern Lakes and Forests (27/27)	This cool, humid area supports extensive forest, with a mix of evergreen and deciduous needleleaf and broadleaf species (Auch and Karstensen, 2015). A variety of wetlands including bogs, marshes, swamps, wet meadows, and open water cover much of the surface.
Mississippi Alluvial Plain and Mississippi Valley Loess Hills (23/37)	This hot, humid landscape has mild winters and originally was covered with bottomland deciduous forests and swamps. Extensive draining and clearing of trees converted the land to agricultural use, which now includes irrigated crops (with some intentional seasonal flooding for growing catfish) and aquaculture (Auch and Karstensen, 2015). The loess plains on the eastern side of the scene have undergone rapid conversions in and out of forests, with the year-round growing conditions enabling landowners to readily turn over land cover in response to changing land management incentives.

Our evaluation goals focused on providing insights into algorithm improvements needed to meet overall LCMAP data quality objectives. The findings presented in this study are based on our efforts to produce land use and cover results that are relatively stable over time and geographic space, are generally consistent between different land cover and use types, and provide consistent results even when input data vary in density and quality.

Figure 1. The red rectangles show the locations of the five Landsat scenes used in our study, with the numeric labels corresponding to path/row designations. The small blue squares are the sample blocks of the Land Cover Trends (LCT) project – the blue blocks located within the red rectangles were used to evaluate the different classification scenarios. The blue lines are the boundaries of Level III Ecoregions.

3 Data

3.1 Landsat data

We used USGS EROS Landsat high-level data products as input to CCDC, including surface reflectance, brightness temperature, and cloud, cloud shadow, and snow masks. The surface reflectance products from Landsats 4–7 were processed with the Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS) (Schmidt et al., 2013) and the surface reflectance products from Landsat 8 were processed with the Landsat 8 Surface Reflectance system (L8SR) (U.S. Geological Survey, 2016). The cloud, cloud shadow, and snow masks were generated by a C program version of the Fmask algorithm that detects

clouds, cloud shadows, and snow using an object-based method (Zhu & Woodcock, 2012; Zhu, Wang, & Woodcock, 2015). We further filtered the dataset for clouds, shadows, and snow with an algorithm called Tmask that uses additional multitemporal information (Zhu & Woodcock, 2014a). We downloaded all available Level 1 Terrain-corrected (L1T) data from Landsats 4–8 from 1982 to 2014 with more than 20% clear (not cloud, cloud shadow, or snow) observations for the five study path/rows. In all, we obtained 2,873 Landsat images, of which 95% were acquired by Landsats 5 and 7 (Table 2; Figure 1).

Table 2. Number of Landsat images used from each Landsat sensor for each path/row.

Number of images	Path/Row				
	46/27	43/34	28/33	27/27	23/27
Landsat 4	5	2	12	1	4
Landsat 5	265	447	356	223	356
Landsat 7	164	299	243	120	250
Landsat 8	21	35	27	13	30
Total Landsat	455	783	638	357	640

3.2 Land Cover Trends (LCT) data

We used LCT data as a basis for training and validation. The LCT project applied a sampling approach to obtain ecoregionally stratified, randomly selected 10 km × 10 km sample blocks (note: a few initial ecoregions had blocks that were 20 km × 20 km) across the conterminous United States (Figure 1) (Loveland et al., 2002). The ecoregions used as strata (U.S. Environmental Protection Agency, 1999, modified from Omernik, 1987) provided spatial units encompassing similar patterns of environmental characteristics and land uses (Omernik, 1987; Gallant et al., 2004). The number of regional sample blocks selected was aimed at meeting the LCT statistical requirements to estimate gross change in land cover with a margin of error of 1% at a 0.85 confidence level (Loveland et al., 2002). The sample blocks selected for the LCT project were manually interpreted based on Landsat imagery, reinforced by consulting high spatial resolution images. The LCT project evaluated various approaches to ensure high

rates of mapping accuracy and concluded that a quality-control approach employing interpreter group review of every block was the most feasible (Auch et al., 2012).

The LCT project originally mapped land cover for five target years (1973, 1980, 1986, 1992, and 2000) spanning the time of Landsat MSS and TM sensors. For the current study we used LCT mapping results from 2000, for which classification was performed at the scale of 30 m pixels, although the minimum mapping unit recognized by analysts was 60 m × 60 m. Therefore, features smaller than the minimum resolution were not represented in the LCT dataset. The LCT thematic scheme recognized 11 classes of land cover/use (Figure 2; Table 3), and the probabilistic sample distribution provided a good range of intra-class variability to be represented, which is important for training data (Pal and Mather, 2003; Rodriguez-Galiano et al., 2012; Cracknell and Reading, 2014). The national coverage by a single data source ensured good consistency in class definitions across our entire study area. As we only were interested in classifying the stable land cover types, the two LCT disturbance classes (*mechanically disturbed* and *non-mechanically disturbed*) were ignored in this study. To test how well our strategy performed across a range of landscapes, LCT data (Soulard et al., 2014) from five path/rows were selected for this analysis (Figure 1). The total number (and percent) of pixels in each land cover type for each path/row is given in Table 4. Note that for some land cover types the total number of potential training pixels in a path/row was likely insufficient for training and validation of land cover; we excluded cover types from the analysis of a given path/row whenever the number of pixels was less than 1,000.

Although the LCT data are considered to be of high quality, they likely are not error free. We performed an assessment of the accuracy following the recommendations of Olofsson et al. (2014) to be certain the data were of sufficient quality. For simple random sampling and targeting overall accuracy as the estimation objective, Cochran (1977, Eq. (4.2)) suggests using a sample size (n) of

$$n = \frac{z^2 O(1-O)}{d^2} \quad \text{Eq. (1)}$$

where,

O: the overall accuracy;

z : the percentile from the standard normal distribution ($z = 1.96$ for a 95% confidence interval);

d : the desired half-width of the confidence interval of O .

By assuming an overall accuracy of 0.95 and a target half-width of the confidence interval of O of 0.025, Eq. (1) was used to estimate the size of a simple random sample of pixels across the LCT blocks within the five study scenes.. This gave a total sample size of 292 pixels, which was rounded to 300. A list frame was developed for all pixels within the LCT sample blocks used for the training analysis across the five path/rows. We excluded from the list any pixels classified as either of the two types of disturbance, which were considered as non-interest categories for this study. We used the SAS software (version 9.3 for Windows) random number generator for a uniform distribution to assign a random number to each pixel, then sorted the pixels in ascending order and extracted the boundaries for the first 300 pixels to overlay in Google earth™, relying on it as our primary source of information for labeling land cover circa 2000. Landsat data and maps from the National Wetlands Inventory (NWI) (Wilén and Bates, 1995; U.S. Fish and Wildlife Service, National Wetlands Inventory, 2016) served as secondary sources of information. Project scientists manually interpreted land cover based on the criteria defined by the LCT project. A primary land cover label was assigned to each pixel, with a secondary label included in cases where mixing of cover types was cause for ambiguity. Results showed the LCT data had an overall rate of accuracy of 91% when only the primary label was considered and 99% when both primary and secondary labels were incorporated. These rates of accuracy confirmed LCT data are suitable as a source of information for developing an optimized strategy for selecting training data.

Figure 2. Land Cover Trends sample blocks located within Landsat path/row 46/27 in Washington, USA. The white rectangle is the boundary of the Landsat scene and the background is high resolution imagery from Google earth™. Note that the two LCT disturbance classes (*mechanically disturbed* and *non-mechanically disturbed*) were ignored for this study.

Table 3. Land cover and land use class descriptions used by the Land Cover Trend project (U.S. Geological Survey, 2015). Note that the two LCT disturbance classes (*mechanically disturbed* and *non-mechanically disturbed*) were ignored for this study.

Class	Description
Water	Areas persistently covered with water, such as streams, canals, lakes, reservoirs, bays, or oceans.
Developed	Areas of intensive use, with much of the land covered with structures (e.g., high density residential, commercial, industrial, transportation, confined livestock operations), or less intensive uses where the land cover matrix includes both vegetation and structures (e.g., low density residential, recreational facilities, cemeteries, etc.), including any land functionally attached to the urban or built-up activity.
Mechanically disturbed	Land in an altered and often non-vegetated state that, due to disturbances by mechanical means, is in transition from one cover type to another. Mechanical disturbances include forest clearcutting, earthmoving, scraping, chaining, reservoir drawdown, and other similar human-induced changes.
Nonmechanically disturbed	Land in an altered and often non-vegetated state that, due to disturbances by nonmechanical means, is in transition from one cover type to another. Nonmechanical disturbances are caused by wind, floods, fire, animals, and other similar phenomenon.
Mining	Areas with extractive mining activities that have a significant surface expression. This includes (to the extent that these features can be detected) mining buildings, quarry pits, overburden, leach, evaporative,

	tailing, or other related components.
Barren	Land composed of natural occurrences of soils, sand, or rocks where less than 10% of the area is vegetated.
Forest	Land where tree-cover density is greater than 10%. Note that cleared forest land (i.e., clearcut logging) will be mapped according to current cover (e.g., disturbed or transitional, shrubland/grassland).
Grass/shrub	Land predominately covered with grasses, forbs, or shrubs. The vegetated cover must cover at least 10% of the area.
Agriculture	Cropland or pastureland in either a vegetated or non-vegetated state used for the production of food and fiber. Note that forest plantations are considered as forests or woodlands regardless of the use of the wood products.
Wetland	Lands where water saturation is the determining factor in soil characteristics, vegetation types, and animal communities. Wetlands are composed of water and vegetated cover.
Ice/snow	Land where the accumulation of snow and ice does not completely melt during the summer period.

Table 4. Number (percent) of training pixels for each land cover category for each path/row. Note that numbers in bold font are those with less than 1000 pixels, which were excluded from analysis in this study.

Land Cover	46/27	43/34	28/33	23/37	27/27
Water	214,667 (5.90%)	14,841 (0.90%)	27,293 (1.01%)	16,819 (1.63%)	84,255 (6.39%)
Developed	510,552 (14.03%)	168,660 (10.28%)	25,665 (0.95%)	10,215 (0.99%)	5,697 (0.43%)
Mining	9,159 (0.25%)	4,662 (0.28%)	1,395 (0.05%)	366 (0.04%)	8,372 (0.64%)
Barren	110,025 (3.02%)	2,756 (0.17%)	705 (0.03%)	516 (0.05%)	65 (0.00%)
Forest	2,154,503 (59.19%)	283,968 (17.31%)	148,862 (5.50%)	325,076 (31.49%)	761,053 (57.74%)
Grass/shrub	225,173 (6.19%)	229,234 (13.97%)	1,178,711 (43.51%)	7,941 (0.77%)	46,318 (3.51%)
Agriculture	266,169 (7.31%)	932,604 (56.84%)	1,316,979 (48.62%)	572,533 (55.47%)	78,622 (5.97%)
Wetland	55,901 (1.54%)	4,048 (0.25%)	9,187 (0.34%)	98,774 (9.5689%)	333,575 (25.31%)
Ice/snow	93,601 (2.57%)	0	0	0	0
Total number of pixels	3,639,750	1,640,773	2,708,797	1,032,240	1,317,957

3.3 Auxiliary data

We tested three groups of auxiliary data, including data used by the National Land Cover Database (NLCD) (Homer and Fry, 2012), Fmask-generated statistics, and CCDC change detection results. These data were available, or could be generated, to cover the conterminous United States.

Six auxiliary variables are used in developing the NLCD, in which five are from Digital Elevation Models (DEM), including elevation, aspect, Compound Topographic Index (CTI) (Moore et al., 1991), slope, and

Positional Index (PI) (<http://nationalmap.gov>; Gesch et al. 2002). The other auxiliary variable, Wetland Potential Index (WPI), was developed by USGS to improve wetland mapping (S. Jin, personal communication). This index was created based on information from the NLCD 2006 map, NWI, and Soil Survey Geographic database (SSURGO) for hydric soils (Natural Resources Conservation Service, 2016). The NWI dataset was compiled from multiple data sources over a time period from mid-1970 to 2011. Similarly, the vintages of source data for SSURGO were from 1980 to 2011. The NLCD land cover data represented land condition circa 2006. The WPI derived by integration all three of these datasets provided a long-term perspective on the likelihood of an area in the conterminous United States being a wetland, rather than reflecting conditions for a specific year. The steady state of the WPI data layer therefore makes it relevant for use with the Landsat images acquired from 1982-2014 for LCMAP classification of land cover for the current study. The WPI represents a categorical ranking index generated according to the convergence of evidence from the above three sources (Figure 3). For example, if none of the three sources indicate the likelihood of a wetland, the WPI will have a value of 0. The value “1” is currently not used in the index, but is reserved for including other potential sources in the future. The information on hydric soils is ranked lowest (a value of “2”) among the three sources, as it references only conditions under which the soil was formed and may not represent current land cover or conditions. The information from the NWI is ranked higher (a value of “3”), as it represents direct existence of a wetland based on manual interpretation of aerial photos. However, the NWI has been derived from different methods and vastly different vintages of data across the country and therefore is less coherent than the wetland mapping results derived from NLCD, which were based on data from Landsat sensors compiled over a much narrower time frame. The NLCD wetland information has thus been given the highest ranking (a value of “4”) in the WPI map. The other categorical rankings have been derived based on how many information sources agree on wetland occurrence and the respective rankings of those sources. The WPI benefits from integrating information from three products and is likely to represent the maximum potential for wetland occurrence, providing a spatial depiction of probable wetland areas.

Figure 3. Wetland Potential Index (WPI) classes and rank values for path/row 43/34. The upper right image is a WPI subset corresponding with the white box in the inset map on the left. The lower right image is a Landsat image of the subset area. NWI: National Wetlands Inventory; NLCD: National Land Cover Database.

Fmask statistics provided information on water, cloud, and snow occurrence for every Landsat image. We estimated the probability of each pixel being covered by water, clouds, or snow based on statistics for percent occurrence in the Fmask results from all historical Landsat imagery (Zhu & Woodcock, 2012). The cloud probability is calculated based on the percent of cloud observations from all available Landsat observations. The snow probability is calculated based on the percent of snow observations from all available Landsat observations that are not classified as cloud or cloud shadow. The water probability is

the percent of water observations from all available clear (not cloud, cloud shadow, or snow) Landsat observations. We included the water probability to assist in classifying water/wetland areas. We included the cloud probability because places with more clouds often signify land cover differences related to less solar radiation and more precipitation. The probability of snow was included because of potential effects on soil wetness that could influence differences in vegetation cover. Moreover, we thought the snow probability variable could increase the classification accuracy for snow.

We used two auxiliary variables derived from the CCDC change detection output: number of changes each Landsat pixel had undergone over the period of record, and the length of time since the previous change. The former variable was hypothesized to be helpful for classifying land covers subject to frequent, cyclical change, such as agriculture, and the latter for distinguishing different transitional cover following a change, such as early succession stages of forest recovery following harvest.

4. Methods

To determine which training data distributions and input features generated the best results, we compared pixels from the land cover classifications generated by CCDC with those from LCT blocks not used for training at pixel level. In essence, we were attempting to optimize a classification process, and the sole criterion used to evaluate alternative training strategies was agreement with the LCT data. Training strategies that improved the agreement between CCDC results and the LCT data were considered the best. This rationale employs a widely used approach of comparing alternative classification results to reference data (Huang et al., 2002; Zhu et al., 2012; Gong et al., 2013). In most of these studies, alternative classification algorithms or image inputs are the primary interest. In our case, alternative training strategies and various sources of ancillary data are the primary focus.

4.1. The CCDC classification method

The CCDC algorithm uses all available Landsat data to estimate time series models, and the models are used for change detection and classification (Zhu & Woodcock, 2014b; Zhu et al., 2015b). The time series

models are composed of harmonic (Fourier) models (Davis, 1986; Rayer, 1971) and a slope component. The coefficients of the time series models determine their shape and contain rich information in the spectral and temporal dimensions (Zhu & Woodcock, 2014b). Three sets of time series models are used, depending on the number of total available clear observations (Zhu et al., 2015b). Instead of classifying the original Landsat images, the coefficients from the time series models and the Root Mean Square Errors (RMSE) calculated during model estimation are used as inputs to the land cover classifier (Zhu & Woodcock, 2014b).

CCDC uses the Random Forest Classifier (RFC) (Breiman, 2001) RFC for land cover classification, as RFC outperformed or at least performed on the same level as other classification techniques (Verikas et al., 2011). The RFC has two input parameters to be adjusted, the number of trees and the number of split variables at each node. In this study, we used the default value for the number of split variables, the square root of the total number of variables. We set the number of trees to 100, a number reported to achieve good classification results while being computationally efficient (Rodriguez-Galiano et al., 2012b). The RFC can also calculate variable importance by estimating the difference in prediction accuracies between the permuted and original out-of-bag samples (i.e., those observations not included in the bootstrap sample). The Random Forest variable importance measurement is more efficient and accurate than univariate screening methods, as it includes not only the influence of each predictor variable separately, but also the multivariate interactions with other predictor variables (Breiman, 2001; Gislason et al., 2006; Chan and Paelinckx, 2008; Archer and Kimes, 2008; Zhu et al., 2012). In this study, we used the RFC-measured importance to determine which auxiliary variables were helpful in improving land cover classification for different locations.

4.2. Accuracy assessment

In this study, the criterion for evaluating training strategies is the agreement between the CCDC-generated land cover classifications and the LCT data. We refer to this agreement as “accuracy” but the

term is used here only in the context of measuring the level of agreement. No sampling, inference or estimation of accuracy of any place or region was attempted or necessary.

The comparison of the CCDC results and LCT data yielded two metrics of performance: overall and minimum accuracy across all land cover classes. Overall accuracy is the proportion of CCDC results that correspond to the LCT data, and minimum accuracy is the minimum of the user's and producer's accuracies, which are the complementary measures of commission and omission errors, respectively, of the LCT data by the CCDC results. When assessing the different training scenarios, the goal was to increase the minimum accuracy without sacrificing the overall accuracy. As the LCT data were used as both training and reference data, we used different LCT blocks for these purposes to avoid optimistic bias caused by spatial autocorrelation in the data (Hammond & Verbyla, 1996; Friedl et al., 2000; Zhen et al., 2013). We used a jackknife technique (Miller, 1974) to select LCT blocks for training and validation. Given n LCT blocks in one Landsat path/row, for each iteration we left out one LCT block for CCDC comparisons and used the remaining $(n-1)$ LCT blocks as the pool for selecting training data. This process was repeated n times until all LCT blocks had been selected for comparison to CCDC results in each Landsat path/row and an overall accuracy and a minimum accuracy were estimated afterward. To test for statistical significance, the process above (n jackknife replications) was repeated 10 times, such that each accuracy rate was estimated 10 times. Paired t-tests were performed on the estimated values for each testing scenarios (comparing a scenario with a previous scenario) to assess if the differences in accuracy of the various testing scenarios were significant at the 5% level. No repetitions were done when we tested the training balance scenarios because of the excessive computation time required.

4.3. Training data extraction methods

Because the LCT dataset contains millions of potential training pixels, four aspects related to selecting training data were examined: distribution, size, balance, and outlier removal. First, we compared outcomes from extracting training data based on equal and proportional selections. Equal selection

provided an equal number of training pixels for each land cover class, whereas proportional selection provided training data per class in proportion to the class occurrence in the LCT data. To ensure that the conclusions of the analysis were not influenced by the total number of training pixels, the number of training pixels was varied from 500 to 25,000 over a series of runs to evaluate resultant classification accuracies for the two distributions across the five path/rows. After determining the preferred distribution, we tested 10 scenarios with different total numbers of training pixels to find the optimum size for the training data set across the five path/rows. To assess the impact of unbalanced training data, the number of training pixels was increased for rare classes and decreased for larger classes. This experiment determined the minimum and maximum number of training pixels per class to achieve both high overall and minimum accuracies. Based on the findings from the previous evaluations, the optimal strategy conditions were adopted to test whether removing outliers from the pool of potential training data further increased the classification accuracy. A pixel was identified as an outlier if the spectral distance from the center of all the training data for a given land cover type, measured by the Mahalanobis distance (Rousseeuw & Van Zomeren, 1990; Hodge & Austin, 2004) and converted to the chi-squared distribution, exceeded thresholds of p -values at 0.005, 0.01, 0.05, and 0.10. In addition, a pixel was identified as an outlier if it was located at (or near) the edge of different land cover classes. The edge pixels were defined by a mathematical morphology operator (Soille, 2013) that eroded the boundaries by 1, 2, 3, 4, and 5 pixels with an eight-connected square operator for all nine LCT classes. Finally, after addressing the aspects related to training data selection, we tested the importance of auxiliary variables in combination with the original CCDC variables. A variable was considered important if its permutation caused more than a 1% increase in error in all classes or more than a 5% increase in error in any individual class, as we deemed these error rates sufficient to impact the final classification results. Overall and minimum accuracies were compared for the “Basic” set (no auxiliary variables), “Selected” set (removal of auxiliary variables that were not important), and “All” set (all auxiliary variables included) to determine if selecting only important variables achieved higher accuracies.

5. Results and discussions

5.1. Impact of training data distribution

The comparison of overall accuracies based on equal versus proportional distributions showed that the latter yielded higher accuracy (approximately 5–15 percentage points improvement) across all sizes of training sets (Figure 4). Improvement in minimum accuracies was favored by the equal distribution approach (Figure 5). The minimum accuracies based on the equal distribution were approximately 1 to 8 percentage points higher than for the proportional distribution, but the difference decreased for two path/rows (27/27 and 46/27) when the number of training pixels increased. This was because the proportional distribution yielded too few training pixels in rare classes for an adequate land cover characterization. As we considered overall accuracy to be more important than minimum accuracy, the proportional distribution was deemed to be the better choice.

Figure 4. Overall accuracies based on equal versus proportional distributions of training pixels across land cover classes for a range of sample sizes from 500 to 25,000. PD: proportional distribution; ED: equal distribution. The parenthetical numbers in the legend indicate the path/row of each Landsat image.

Figure 5. Minimum accuracies based on equal versus proportional distributions of training pixels across land cover classes for a range of sample sizes from 500 to 25,000. PD: proportional distribution; ED: equal distribution. The parenthetical numbers in the legend indicate the path/row of each Landsat image.

5.2. Impact of training data size

The effect of the size of the training data set was examined by varying the number of pixels from 500 to 25,000 with proportional distribution (Figures 6–7). Scenarios were compared by paired *t*-tests of the estimated accuracies. In all five path/rows, significant (at the 5% level) increases in overall accuracy were gained when using 500 to 5,000 training pixels. Improvements in accuracy were less significant when the number of training pixels exceeded 5,000. In two path/rows (27/27 and 46/27) a slight upward trend was observed when the total number of training pixels exceeded 10,000, whereas this increasing trend plateaued (less than 0.1 percentage point increase in overall accuracy) when the sample size reached

20,000. For the minimum accuracies, two path/rows increased linearly with increased number of training pixels and three path/rows were quite stable, with minimum accuracies close to zero. Though larger training sets can increase minimum accuracy, the improvements were insignificant (at the 5% level) when the set exceeded 20,000 pixels. A total of 20,000 training pixels was therefore determined to be an optimum size.

Figure 6. Overall accuracies based on proportional distributions of training pixels across land cover classes for a total of 500, 1,000, 2,500, 5,000, 7,500, 10,000, 12,500, 15,000, 20,000, and 25,000 training pixels. Note that the black circles indicate scenarios with accuracies that are significantly (at the 5% level) higher than the previous scenarios. PD: proportional distribution. The parenthetical numbers in the legend indicate the path/row of each Landsat image.

Figure 7. Minimum accuracies based on proportional distributions of training pixels across land cover classes for a total of 500, 1,000, 2,500, 5,000, 7,500, 10,000, 12,500, 15,000, 20,000, and 25,000 training pixels. Note that the black circles indicate scenarios with accuracies that are significantly (at the 5% level) higher than the previous scenarios. ED: equal distribution. The parenthetical numbers in the legend indicate the path/row of each Landsat image.

5.3. Impact of training data balance

Proportional distribution with a total of 20,000 training pixels yielded higher overall accuracy, but lower minimum accuracy. For rare classes, very few training pixels will be extracted, and this unbalanced data set will greatly reduce the accuracy of these classes. To achieve a better balance, we first generated training pixels based on proportional distribution with a total of 20,000 training pixels, and if the number of training pixels for a class was less than the minimum number, we increased the number of pixels to the minimum number. Conversely, if the number of training pixels for a class was more than the maximum

number, we reduced the number of pixels to the maximum number. The minimum number of training pixels per class was increased from 0 to 2,000 at increments of 200 and the maximum number of training pixels per class was decreased from unlimited to 2,000. Overall and minimum accuracies for a total of 66 combinations for all five path/rows. As there are five path/rows, for each combination there are five accuracies. Figure 8 shows the mean overall accuracies (five path/rows) calculated for each combination. The lower left corner of the figure shows the highest mean overall accuracy and the upper right corner shows the lowest. The diagonal track from the lower left to the upper right of the figure represents a transition from proportional to equal distribution of training pixels. The highest mean overall accuracy (83.08%) was located at the lower left corner, with no minimum and maximum constraints. The lowest average overall accuracy (74.83%) was located at the upper right corner, with number of training data for all classes equal to 2,000. At the lower left region, the mean overall accuracies were quite similar to one another (dark red colors in Figure 8).

Figure 9 shows the mean minimum accuracies (five path/rows) calculated for each combination. Unlike Figure 8, Figure 9 reveals no spatial trend in the matrix of minimum accuracies, but the relatively higher accuracies are located in a region with minimum number of training pixels between 400 and 800 and maximum number of training pixels between 2,000 and 8,000 (darker red colors in Figure 9). As we favored overall accuracy over minimum accuracy, we looked for combinations where overall accuracy was not significantly lower (5% level) than the combination providing the highest overall accuracy while simultaneously achieving higher minimum accuracy. We performed a paired t-test for the overall accuracies derived from all the 66 combinations against the combination with highest overall accuracy. Combinations demarked with white rectangles in Figure 8 represent levels of accuracy not significantly different (at the 5% level) from the best accuracy. For these 11 combinations, the highest mean minimum accuracy (8.17%) was obtained with a maximum number of 8,000 training pixels and a minimum number of 600 training pixels per class (Figures 8–9). For a good strategy for classifying areas approximately the

size of a Landsat path/row with legends of 9 classes, we therefore recommend a minimum of 600 and a maximum of 8,000 training pixels per class.

Figure 8. Overall accuracies with minimum number of training pixels per class increasing from 0 to 2000 and maximum number of training pixels per class decreasing from 10,000 to 2,000. White rectangles represent levels of accuracy not significantly different (at the 5% level) from the best accuracy.

Figure 9. Minimum accuracies with minimum number of training pixels per class increasing from 200 to 2000 and maximum number of training pixels per class decreasing from 10,000 to 2,000.

5.4. Impact of outliers

Figures 10–11 show the overall and minimum accuracies at p -values of 0.005, 0.01, 0.05, and 0.10 after outliers were removed in the spectral dimension (same variables used in CCDC classification). For overall accuracy, there were two significant increases (at the 5% level) for a single path/row (23/37) when outliers were removed at p -values of 0.005 and 0.05; however, the actual magnitude of improvement was minor. Moreover, the overall accuracies for the other four path/rows decreased with increasing p -values. For minimum accuracy, only one path/row (27/27) showed a significant (at the 5% level) increase in accuracy at a p -value of 0.005; the other four sites exhibited either decreased or nearly level rates of accuracy. These results based on the measured distance in the spectral dimension provided no rationale for removal of outliers from the pool of potential training data.

Figures 12–13 show the overall and minimum accuracies after outliers were removed by eroding 1, 2, 3, 4, and 5 edge pixels in the spatial dimension. Only two cases (1-pixel erosion for 28/33 and 3-pixel erosion for 43/34) yielded significant increases in accuracy, but the magnitude of improvement was minor. Generally, the overall accuracies for all five path/rows were either constant or decreased as the number of eroded pixels increased. For minimum accuracy, two path/rows (28/33 and 46/27) achieved optimum accuracy with 1 and 3 pixels eroded, respectively, but the other three sites had either decreasing or nearly constant accuracy with increasing removal of outliers. Therefore, we found little or no support for removal of outliers based on the the measured distance in the spatial dimension.

An explanation for the limited impact of outlier removal in the training data is that the RFC is relatively robust to noise. Rodriguez-Galiano et al. (2012b) reported that as long as a noise threshold of 20% is not exceeded, the error rate for the RFC overall accuracy remains almost the same. Another possible explanation is that removing pixels that are spectrally or spatially different simultaneously removes most of the mixed pixels, which are reported to be more useful than pure pixels in machine learning classifiers (Foody & Mathur, 2006). Therefore, we do not recommend outlier removal, as it had only a minor (and inconsistent) effect on accuracy.

Figure 10. Overall accuracies when outliers (measured in the spectral dimension) were removed at p -values of 0.005, 0.01, 0.05, and 0.10. Note that the black circles indicate scenarios with accuracies that were significantly higher (at the 5% level) than the previous scenarios. The numbers in the legend indicate the path/row of each Landsat image.

Figure 11. Minimum accuracies at p -values of 0.005, 0.01, 0.05, and 0.10 after outliers (measured in the spectral dimension) were removed. Note that the black circles indicate scenarios with accuracies that were significantly higher (at the 5% level) than the previous scenarios. The numbers in the legend indicate the path/row of each Landsat image.

Figure 12. Overall accuracies after outliers (measured in the spatial dimension) were removed by eroding 1, 2, 3, 4, and 5 edge pixels. Note that the black circles indicate the scenarios with accuracies that were significantly higher (at the 5% level) than the previous scenarios. The numbers in the legend indicate the path/row of each Landsat image.

Figure 13. Minimum accuracies after outliers (measured in the spatial dimension) were removed by eroding 1, 2, 3, 4, and 5 edge pixels. Note that the black circles indicate the scenarios with accuracies that were significantly higher (at the 5% level) than the previous scenarios. The numbers in the legend indicate the path/row of each Landsat image.

5.5. Impact of auxiliary data

Building on our findings thus far, we tested the use of 11 auxiliary data layers to further improve classification accuracy. Figure 14 shows the variable importance, evaluated by mean increase in error for all land cover classes combined and individually when each auxiliary layer was permuted. The mean increase in error was computed for all five path/rows, and the values shown in Figure 14 were those with the largest magnitude. We considered a variable to be important if permuting it could cause a mean

increase in error of more than 1 percentage point for all classes combined or more than 5 percentage points for an individual class for any path/row. Three variables (CTI, number of changes, and length of time since the previous change) did not meet any of the criteria, so we would not include them as part of our recommended strategy. The CTI variable was anticipated to help find potential places in the landscape where water could collect to improve mapping of wetlands and water bodies. The WPI variable turned out to be much more useful for identifying these features, which were well represented in path/rows 27/27 and 23/37. We included the variable for the number of changes a pixel underwent to help identify high turnover land cover in cropland areas. Path/rows 23/37 and 43/34 provided extensive cropland with high turnover frequency to test this variable, but we found that the RMSE was a much stronger indicator for these agricultural lands. We had hoped that the variable representing the time since the last change would be helpful for identifying early successional changes following a vegetation disturbance. Path/row 46/27 provided a lot of post-harvested forest patches to test this variable, but it did not turn out to be informative for this or any other type of land cover change we mapped.

The most helpful auxiliary variable was elevation (DEM), especially for forest classification (Figure 14). Some auxiliary data may only have substantial impacts on certain land cover classes. For example, the WPI, snow probability, and water probability variables had large impacts on wetland, snow/ice, and water classes, respectively. We tested three scenarios to better determine that adding auxiliary data could improve classification accuracy and removing the less important layers would not have any impact: 1) “Basic” scenario – using only the basic CCDC-derived variables; 2) “Selected” scenario - adding only the important auxiliary variables; and 3) “All” auxiliary scenario - adding all auxiliary variables (Figures 15–16). Results from the Selected scenario were significantly higher (at the 5% level) than for the Basic scenario in four of five path/rows, with two path/rows having more than 1 percentage point increase in overall accuracy. The overall accuracies between the Selected and All auxiliary scenarios were very similar. Although three of five path/rows showed significant increases (at the 5% level), the improvement in accuracy was small. For minimum accuracy, two of five path/rows showed significant increases (at the

5% level) in the Selected and All auxiliary scenarios, but the improvements were smaller. Weighing the benefits of the potential to improve the classification results, the magnitude of the improvement, and the cost of preparing, storing, and processing auxiliary data, we concluded that the Selected scenario best served the needs of the LCMAP goal for continuous monitoring and characterization of land change.

Figure 14. Variable importance for 11 auxiliary data layers measured by mean increase in classification error. The red dashed lines indicate thresholds of 1% and 5% of mean increases in error for all classes and for individual classes, respectively. cti: compound topographic index; dem: digital elevation model; pi: positional index; wpi: wetland potential index; water prob: water probability; snow prob: snow probability; cloud prob: cloud probability; number: number of changes that have occurred; length: length of time since the previous change.

Figure 15. Overall accuracies for Basic, Selected, and All auxiliary layer scenarios. Note that the black circles indicate scenarios with accuracies that were significantly higher (at the 5% level) than the previous scenarios. The numbers in the legend indicate the path/row of each Landsat image.

Figure 16. Minimum accuracies for Basic, Selected, and All auxiliary layer scenarios. Note that the black circles indicate the scenarios with accuracies that were significantly higher (at the 5% level) than the previous scenarios. The numbers in the legend indicate the path/row of each Landsat image.

5.6. Accuracies based on the optimized strategy

Based on the findings, we identified a strategy for selection of training and auxiliary data and applied it to all LCT blocks in the study area; Table 5 shows the producer's, user's and overall accuracies for all land cover types. Three of the minimum accuracies were from the mining class and two were from the grass/shrub class, all of which were small classes, making them harder to classify. Mining is a land use, rather than a land cover, further complicating the mapping. The developed, barren, and wetland classes also occupied relatively small proportions of the landscape and had low accuracies compared to better represented land covers (e.g., water, forest, and agriculture). The overall accuracy ranged from 80 to 88%

when applying the optimal strategy. In comparison, the original strategy of selecting training data for use with CCDC did not incorporate auxiliary data and included approximately 500 pixels per class (Zhu & Woodcock, 2014b); when applied to the five path/rows the original method resulted in mean overall and minimum accuracies of 69% and 4%, respectively (see Figures 4–5). Pursuing the optimal strategy determined from the current research achieved 84% in mean overall accuracy (15 percentage points increase) and 8% (4 percentage points increase) in mean minimum accuracy (Table 5).

Table 5. Producer’s (P), user’s (U), overall, and minimum accuracies of the CCDC-generated maps when assessed against the Land Cover Trends data.

<i>Land Cover</i>	<i>46/27 P</i>	<i>46/27 U</i>	<i>43/34 P</i>	<i>43/34 U</i>	<i>28/33 P</i>	<i>28/33 U</i>	<i>23/37 P</i>	<i>23/37 U</i>	<i>27/27 P</i>	<i>27/27 U</i>
<i>Water</i>	87.29	92.72	69.71	59.17	75.59	76.94	61.71	75.62	91.06	92.49
<i>Developed</i>	70.9	68.42	74.22	84.8	51.63	62.85	23.96	32.75	35.51	22.5
<i>Mining</i>	44.14	16.02	0	0	61.96	17.73	NA	NA	8.39	37.23
<i>Barren</i>	42.88	54.94	60.24	30.73	NA	NA	NA	NA	NA	NA
<i>Forest</i>	91.16	87.89	85.11	85.48	60.08	66.66	90.32	78.97	86.04	85.48
<i>Grass/shrub</i>	36.17	45.59	73	73.88	89.58	82.12	0.15	0.16	4.47	29.55
<i>Agriculture</i>	69.37	75.96	96.5	94.43	82.69	89.09	90.87	91.4	73.13	67.48
<i>Wetland</i>	30.68	48.53	30.26	18.53	34.07	24.23	51.64	82.4	77.69	70.84
<i>Ice/snow</i>	87.14	77.41	NA	NA	NA	NA	NA	NA	NA	NA
<i>Overall</i>	80.49		88.21		83.91		85.1		79.9	
<i>Minimum</i>	16.02		0		17.73		0.14		4.47	

We note that the actual accuracy of the CCDC algorithm is likely to be higher than the accuracies provided in Table 4. One reason is that the “accuracy” computed here is actually the agreement between CCDC and the LCT dataset. Most concerning in this regard is the different level of generalization in the two datasets. The LCT had a 60 m minimum mapping unit and was derived using visual interpretation, leading to a higher level of generalization than produced by an automated, per-pixel classification process operating at 30 m resolution. Figure 17 illustrates the kinds of differences that can occur between the LCT and CCDC classification results. For location A, small areas of buildings and roads were included in the Agriculture class in the LCT data, but mapped as Developed by CCDC. And, as with all datasets, there are sometimes errors in the LCT data; close examination of the high resolution imagery for location B shows that areas of agricultural use were incorrectly labeled as Developed in LCT data. Given that LCT

625 data were used as reference data for computing the “accuracy” of the CCDC map results, the true overall
 626 accuracies may be underestimated.

627
 628 Figure 17. The upper-right image shows one 10 km × 10 km classified Land Cover Trends block from 46/27. The other three
 629 images are from the same zoomed area of this image (the black rectangle). The upper-left image shows high-resolution imagery

from January 6, 2002, in Google earth™. The lower-left image shows the Land Cover Trends (LCT) map for the zoomed-in extent. The lower-right image shows the corresponding CCDC classification results. Hatch lines in the lower two images are areas of disagreement between LCT and CCDC results.

6. Conclusions

Several conclusions can be drawn from this study: 1) Training data should be extracted based on proportional occurrence of land cover classes, with a total of 20,000 pixels distributed across classes for a mapping area with similar dimensions to a Landsat path/row and a legend scheme of about 10 classes. 2) We recommend that training data are balanced among larger and smaller classes by using a minimum of 600 and a maximum of 8,000 training pixels. 3) Removal of outlier pixels from the potential training pool based on spectral or spatial distances is not necessary. 4) Eight auxiliary variables helped improve our classification accuracy; five (aspect, elevation, positional index, slope, and Wetland Potential Index) are used in developing the NLCD and three are from Fmask statistics (water probability, snow probability, and cloud probability). The strategy we developed for selecting training data and auxiliary variables was designed for operational automated land cover mapping of the conterminous United States for the LCMAP program. A core classification strategy that includes accurate and effectively used training data is a prerequisite for meeting the LCMAP objective to characterize land cover and land change at any point in the Landsat record and to detect and characterize change as it occurs. The strategy described here is a part of that foundation. We recognize that this strategy may be modified if other classifiers, inputs, or land cover legends are used, but the core methodology presented in this paper for optimizing the strategy can still be applied.

Acknowledgments

This work was supported by the USGS LandCarbon, Land Remote Sensing, and Land Change Science programs, partially under USGS contracts G13PC00028 (Z. Zhu and S. Jin), G15PC00012 (B. Pengra and D. Dahal), and G15PC00012 (L. Yang). Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References:

- Anderson, J.R., Hardy, E.E., Roach, J.T., & Witmer, R.E. A land use and land cover classification system for use with remote sensor data. U.S. Government Printing Office: 1976; Vol. 964.
- Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52(4), 2249-2260.
- Auch, R. F. , Drummond, J. A. , Sayler, K. L., Gallant, A. L., & Acevedo, W. (2012). An approach to assess land-cover trends in the conterminous United States (1973–2000). In C. Giri (Ed.), *Remote sensing of land use and land cover — principles and applications* (pp. 351–368). Boca Raton: CRC Press.
- Auch, R.F., and Karstensen, K.A., eds., 2015, Status and trends of land change in the Midwest–South Central United States—1973 to 2000: U.S. Geological Survey Professional Paper 1794–C, 190 p.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32.
- Chan, J. C. W., & Paelinckx, D. (2008). Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment*, 112(6), 2999-3011.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
- Chen, J., Chen, J., Liao, A., Cao, X., Chen, L., Chen, X., et al. (2015). Global land cover mapping at 30m resolution: A POK-based operational approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 103, 7-27.

675 Corcoran, J. M., Knight, J. F., & Gallant, A. L. (2013). Influence of multi-source and multi-temporal
676 remotely sensed and ancillary data on the accuracy of random forest classification of wetlands in
677 Northern Minnesota. *Remote Sensing*, 5(7), 3212-3238.

678 Cochran, W. G. (1977). *Sampling techniques*. John Wiley & Sons. Cracknell, M.J., and Reading, A.M.
679 (2014). Geological mapping using remote sensing data: A comparison of five machine learning
680 algorithms, their response to variations in the spatial distribution of training data and the use of
681 explicit spatial information: *Computers & Geosciences*, v. 63, p. 22-33.

682 Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007).
683 Random forests for classification in ecology. *Ecology*, 88(11), 2783-2792.

684 Davis, J. C. (1986). *Statistics and Data Analysis in Geology*, Second Edition, *J. Wiley and Sons*, New
685 York, N. Y., 646.

686 Estabrooks, A., Jo, T., & Japkowicz, N. (2004). A multiple resampling method for learning from
687 imbalanced data sets. *Computational Intelligence*, 20(1), 18-36.

688 Foody, G. M. (1995). Land cover classification by an artificial neural network with ancillary
689 information. *International Journal of Geographical Information Systems*, 9(5), 527-542.

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

692 Foody, G. M., & Arora, M. K. (1997). An evaluation of some factors affecting the accuracy of
693 classification by an artificial neural network. *International Journal of Remote Sensing*, 18(4), 799-
694 810.

695 Foody, G. M., & Mathur, A. (2006). The use of small training sets containing mixed pixels for accurate
696 hard image classification: Training on mixed spectral responses for classification by a SVM. *Remote*
697 *Sensing of Environment*, 103(2), 179-189.

698 Freeman, E. A., Moisen, G. G., & Frescino, T. S. (2012). Evaluating effectiveness of down-sampling for
699 stratified designs and unbalanced prevalence in Random Forest models of tree species distributions
700 in Nevada. *Ecological Modelling*, 233, 1-10.

701 Friedl, M. A., McIver, D. K., Hodges, J. C., Zhang, X. Y., Muchoney, D., Strahler, A. H., et al. (2002).
702 Global land cover mapping from MODIS: algorithms and early results. *Remote Sensing of*
703 *Environment*, 83(1), 287-302.

704 Friedl, M. A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A., & Huang, X.
705 (2010). MODIS Collection 5 global land cover: Algorithm refinements and characterization of new
706 datasets. *Remote Sensing of Environment*, 114(1), 168-182.

707 Friedl, M. A., Woodcock, C., Gopal, S., Muchoney, D., Strahler, A. H., & Barker-Schaaf, C. (2000). A
708 note on procedures used for accuracy assessment in land cover maps derived from AVHRR data.

709 Fry, J.A., Xian, G., Jin, S., Dewitz, J.A., Homer, C.G., LIMIN, Y., Barnes, C.A., Herold, N.D. and
710 Wickham, J.D., (2011). Completion of the 2006 national land cover database for the conterminous
711 United States. *Photogrammetric engineering and remote sensing*, 77(9), pp.858-864.

712 Gallant, A. L. (2015). The Challenges of Remote Monitoring of Wetlands. *Remote Sensing*, 7(8), 10938-
713 10950.

714 Gallant, A.L., Loveland, T.R., Sohl, T.L., and Napton, D.E. (2004). Using an ecoregion framework to
715 analyze land-cover and land-use dynamics: *Environmental Management*, v. 34, no. 1, p. S89-S110.

716 Gesch, D.B., Oimoen, M., Greenlee, S.K., Nelson, C.A., Steuck, M., and Tyler, D., 2002, The National
 717 Elevation Dataset: Photogrammetric Engineering & Remote Sensing, v. 68, no. 1, p. 5–11.

718 Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2006). Random forests for land cover
 719 classification. *Pattern Recognition Letters*, 27(4), 294-300.

720 Gong, P., Wang, J., Yu, L., Zhao, Y., Zhao, Y., Liang, L., et al. (2013). Finer resolution observation and
 721 monitoring of global land cover: First mapping results with Landsat TM and ETM+
 722 data. *International Journal of Remote Sensing*, 34(7), 2607-2654.

723 Gopal, S., Woodcock, C. E., & Strahler, A. (1999). Fuzzy neural classification of global land cover from
 724 a 1 degree AVHRR Data Set. *Remote Sensing of Environment*, 67, 230-243.

725 Ham, J., Chen, Y., Crawford, M. M., & Ghosh, J. (2005). Investigation of the random forest framework
 726 for classification of hyperspectral data. *Geoscience and Remote Sensing, IEEE Transactions*
 727 *on*, 43(3), 492-501.

728 Hammond, T. O., & Verbyla, D. L. (1996). Optimistic bias in classification accuracy
 729 assessment. *International Journal of Remote Sensing*, 17(6), 1261-1266.

730 Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence*
 731 *Review*, 22(2), 85-126.

732 Homer, C., Dewitz, J., Fry, J., Coan, M., Hossain, N., Larson, C., Herold, N., McKerrow, A., VanDriel,
 733 J.N. and Wickham, J., (2007). Completion of the 2001 national land cover database for the
 734 conterminous United States. *Photogrammetric Engineering and Remote Sensing*, 73(4), p.337.

735 Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J., Herold, N., Wickham, J.
 736 and Megown, K., (2015). Completion of the 2011 National Land Cover Database for the

737 conterminous United States—representing a decade of land cover change information.
 738 Photogrammetric Engineering & Remote Sensing, 81(5), pp.345-354.

739 Homer, C. and Fry, J., 2012, The National Land Cover Database: U.S. Geological Survey Fact Sheet
 740 2012-3020.

741 Huang, C., Davis, L. S., & Townshend, J. R. (2002). An assessment of support vector machines for land
 742 cover classification. *International Journal of Remote Sensing*, 23(4), 725-749.

743 Jin, H., Stehman, S. V., & Mountrakis, G. (2014). Assessing the impact of training sample selection on
 744 accuracy of an urban classification: a case study in Denver, Colorado. *International Journal of*
 745 *Remote Sensing*, 35(6), 2067-2081.

746 Lawrence, R. L., Wood, S. D., & Sheley, R. L. (2006). Mapping invasive plants using hyperspectral
 747 imagery and Breiman Cutler classifications (RandomForest). *Remote Sensing of*
 748 *Environment*, 100(3), 356-362.

749 Ling, C. X., & Li, C. (1998, August). Data Mining for Direct Marketing: Problems and Solutions.
 750 In *KDD* (Vol. 98, pp. 73-79).

751 Loveland, T. R., Reed, B. C., Brown, J. F., Ohlen, D. O., Zhu, Z., Yang, L. W. M. J., & Merchant, J. W.
 752 (2000). Development of a global land cover characteristics database and IGBP DISCover from 1 km
 753 AVHRR data. *International Journal of Remote Sensing*, 21(6-7), 1303-1330.

754 Loveland, T., Sohl, T., Stehman, S., Gallant, A., Sayler, K., Napton, D. (2002). A strategy for estimating
 755 the rates of recent united states land-cover changes. *Photogrammetric Engineering & Remote*
 756 *Sensing*, 68, 1091-1099.

757 Lu, D., & Weng, Q. (2007). A survey of image classification methods and techniques for improving
 758 classification performance. *International Journal of Remote Sensing*, 28(5), 823-870

759 Maselli, F., Conese, C., De Filippis, T., & Romani, M. (1995). Integration of ancillary data into a
760 maximum-likelihood classifier with nonparametric priors. *ISPRS Journal of Photogrammetry and*
761 *Remote Sensing*, 50(2), 2-11.

762 Mather, P. M. (2004). *Computer Processing of Remotely-Sensed Images*, 3rd edition. Chichester: Wiley

763 Mellor, A., Boukir, S., Haywood, A., & Jones, S. (2015). Exploring issues of training data imbalance and
764 mislabelling on random forest performance for large area land cover classification using the
765 ensemble margin. *ISPRS Journal of Photogrammetry and Remote Sensing*, 105, 155-168.

766 Millard, K., & Richardson, M. (2015). On the importance of training data sample selection in random
767 forest image classification: A case study in peatland ecosystem mapping. *Remote Sensing*, 7(7),
768 8489-8515.

769 Miller, R. G. (1974). The jackknife-a review. *Biometrika*, 61(1), 1-15.

770 Moore, I.D., Grayson, R.B. and Ladson, A.R. 1991. Digital Terrain Modelling: A Review of
771 Hydrological, Geomorphological, and Biological Applications. *Hydrological Processes*, 5:3-30.

772 Natural Resources Conservation Service, U.S. Department of Agriculture, 2016. Web Soil Survey.
773 Available at <http://websoilsurvey.nrcs.usda.gov/>. Accessed 8.03.16.

774 Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., & Wulder, M. A. (2014).
775 Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of*
776 *Environment*, 148, 42-57.

777 Olofsson, P., Foody, G. M., Stehman, S. V., & Woodcock, C. E. (2013). Making better use of accuracy
778 data in land change studies: Estimating accuracy and area and quantifying uncertainty using
779 stratified estimation. *Remote Sensing of Environment*, 129, 122-131.

780 Omernik, J.M. (1987). Ecoregions of the conterminous United States, *Annals of the Association of*
781 *American Geographers*, 77:118-125.

782 Ozdogan, M., & Gutman, G. (2008). A new methodology to map irrigated areas using multi-temporal
783 MODIS and ancillary data: An application example in the continental US. *Remote Sensing of*
784 *Environment*, 112(9), 3520-3537.

785 Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote*
786 *Sensing*, 26(1), 217-222.

787 Pal, M., and Mather, P.M. (2003), An assessment of the effectiveness of decision tree methods for land
788 cover classification: *Remote Sensing of Environment*, v. 86, no. 4, p. 554-565.

789 Piper, J. (1992). Variability and bias in experimentally measured classifier error rates. *Pattern*
790 *Recognition Letters*, 13(10), 685-692.

791 Radoux, J., Lamarche, C., Van Bogaert, E., Bontemps, S., Brockmann, C., & Defourny, P. (2014).
792 Automated training sample extraction for global land cover mapping. *Remote Sensing*, 6(5), 3965-
793 3987.

794 Rayner, J. N. (1971). *An Introduction to Spectral Analysis*, *Pion Ltd.*, London, 174.

795 Rodriguez-Galiano, V. F., Chica-Olmo, M., Abarca-Hernandez, F., Atkinson, P. M., & Jeganathan, C.
796 (2012a). Random Forest classification of Mediterranean land cover using multi-seasonal imagery
797 and multi-seasonal texture. *Remote Sensing of Environment*, 121, 93-107.

798 Rindfuss, R.R., Walsh, S.J., Turner, B.L., Fox, J. and Mishra, V., (2004). Developing a science of land
799 change: challenges and methodological issues. *Proceedings of the National Academy of Sciences of*
800 *the United States of America*, 101(39), pp.13976-13981.

801 Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M., and Rigol-Sanchez, J.P. (2012b), An
802 assessment of the effectiveness of a random forest classifier for land-cover classification: ISPRS
803 Journal of Photogrammetry and Remote Sensing, v. 67, p. 93-104.

804 Rogan, J., Miller, J., Stow, D., Franklin, J., Levien, L., & Fischer, C. (2003). Land-cover change
805 monitoring with classification trees using Landsat TM and ancillary data. Photogrammetric
806 Engineering & Remote Sensing, 69(7), 793-804.

807 Rousseeuw, P. J., & Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage
808 points. Journal of the American Statistical Association, 85 (411), 633-639.

809 Schmidt, G., Jenkerson, C., Masek, J., Vermote, E., & Gao, F. (2013). Landsat ecosystem disturbance
810 adaptive processing system (LEDAPS) algorithm description (No. 2013-1057). U.S. Geological
811 Survey.

812 Sleeter, B. M., & Wilson, T. (Eds.). (2012). Status and Trends of Land Change in the Western United
813 States--1973 to 2000. U.S. Department of the Interior, U.S. Geological Survey.

814 Sohl, T.L., Gallant, A.L., and Loveland, T.R., 2004, Land surface change and implications for project
815 design: Photogrammetric Engineering and Remote Sensing, v. 70, No. 4, p. 439-448.

816 Soille, P. (2013). Morphological image analysis: principles and applications. Springer Science &
817 Business Media.

818 Soulard, C.E., Acevedo, W., Auch, R.F., Sohl, T.L., Drummond, M.A., Sleeter, B.M., Sorenson, D.G.,
819 Kambly, S., Wilson, T.S., Taylor, J., Sayler, K.L., Stier, M.P., Barnes, C., Methven, S.C., Loveland,
820 T.R., Headley, R., and Brooks, M.S., 2014, Land cover trends dataset, 1973–2000: U.S. Geological
821 Survey Data Series 844, 10 p. Available at <http://dx.doi.org/10.3133/ds844>.

822 Stehman, S. V., & Foody, G. M. (2009). Accuracy assessment. The SAGE handbook of remote sensing,
823 297-309. Stehman, S. V., Sohl, T. L., & Loveland, T. R. (2003). Statistical sampling to characterize
824 recent United States land-cover change. *Remote Sensing of Environment*, 86(4), 517-529.

825 Strahler, A. H. (1980). The use of prior probabilities in maximum likelihood classification of remote
826 sensing data. *Remote Sensing of Environment*, 10, 135-163.

827 Stuckens, J., Coppin, P. R., & Bauer, M. E. (2000). Integrating contextual information with per-pixel
828 classification for improved land cover classification. *Remote Sensing of Environment*, 71(3), 282-
829 296.

830 Taylor, J.L., Acevedo, W., Auch, R.F., and Drummond, M.A., eds., (2015). Status and trends of land
831 change in the Great Plains of the United States—1973 to 2000: U.S. Geological Survey Professional
832 Paper 1794–B, 180 p.

833 Turner, B.L., Lambin, E.F. and Reenberg, A., (2007). The emergence of land change science for global
834 environmental change and sustainability. *Proceedings of the National Academy of Sciences*,
835 104(52), pp.20666-20671

836 U.S. Environmental Protection Agency (1999). Level 111 Ecoregions of the Continental United States,
837 U.S. Environmental Protection Agency, National Health and Environmental Effects Research
838 Laboratory, Corvallis, Oregon (1:7,500,000-scale map).

839 U.S. Fish and Wildlife Service, National Wetlands Inventory home page: <http://www.fws.gov/wetlands>
840 (accessed March 16, 2016).

841 U.S. Geological Survey. (2016). Landsat 8 Product Guide, Version 2.0. Available at
842 http://landsat.usgs.gov/documents/provisional_l8sr_product_guide.pdf (accessed March 16, 2016).

843 U.S. Geological Survey. (2015). Land Cover Trends Project. Available at
844 <http://landcovertrends.usgs.gov/main/classification.html> (accessed March 16, 2016).

845 Verikas, Antanas, Adas Gelzinis, and Marija Bacauskiene. "Mining data with random forests: A survey
846 and results of new tests." *Pattern Recognition* 44, no. 2 (2011): 330-349.

847 Vogelmann, J.E., Howard, S.M., Yang, L., Larson, C.R., Wylie, B.K. and Van Driel, N., 2001.
848 Completion of the 1990s National Land Cover Data Set for the conterminous United States from
849 Landsat Thematic Mapper data and ancillary data sources. *Photogrammetric Engineering and*
850 *Remote Sensing*, 67(6).

851 Vogelmann, J. E., Sohl, T. L., Campbell, P. V., & Shaw, D. M. (1998). Regional land cover
852 characterization using Landsat Thematic Mapper data and ancillary data sources. *Environmental*
853 *Monitoring and Assessment*, 51(1-2), 415-428.

854 Weiss, G. M., & Provost, F. (2003). Learning when training data are costly: the effect of class distribution
855 on tree induction. *Journal of Artificial Intelligence Research*, 315-354.

856 Wilen, B.O., & Bates, M.K. The U.S. Fish and Wildlife Service's National Wetlands Inventory project.
857 *Veg* 1995, 118, 153-169.

858 Wright, C., & Gallant, A. (2007). Improved wetland remote sensing in Yellowstone National Park using
859 classification trees to combine TM imagery and ancillary environmental data. *Remote Sensing of*
860 *Environment*, 107(4), 582-605.

861 Wulder, M. A., White, J. C., Goward, S. N., Masek, J. G., Irons, J. R., Herold, M., et al. (2008). Landsat
862 continuity: Issues and opportunities for land cover monitoring. *Remote Sensing of*
863 *Environment*, 112(3), 955-969.

864 Wulder, M.A., White, J.C., Loveland, T.R., Woodcock, C.E., Belward, A.S., Cohen, W.B., Fosnight,
865 E.A., Shaw, J., Masek, J.G. and Roy, D.P., (2015). The global Landsat archive: Status,
866 consolidation, and direction. *Remote Sensing of Environment*.

867 Yan, L., & Roy, D. P. (2015). Improved time series land cover classification by missing-observation-
868 adaptive nonlinear dimensionality reduction. *Remote Sensing of Environment*, 158, 478-491.

869 Zhen, Z., Quackenbush, L. J., Stehman, S. V., & Zhang, L. (2013). Impact of training and validation
870 sample selection on classification accuracy and accuracy assessment when using reference polygons
871 in object-based classification. *International journal of remote sensing*, 34(19), 6914-6930.

872 Zhu, Z., Wang, S., & Woodcock, C. E. (2015a). Improvement and expansion of the Fmask algorithm:
873 cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images. *Remote*
874 *Sensing of Environment*, 159, 269-277.

875 Zhu, Z., & Woodcock, C. E. (2012). Object-based cloud and cloud shadow detection in Landsat
876 imagery. *Remote Sensing of Environment*, 118, 83-94.

877 Zhu, Z., & Woodcock, C. E. (2014a). Automated cloud, cloud shadow, and snow detection in
878 multitemporal Landsat data: An algorithm designed specifically for monitoring land cover change.
879 *Remote Sensing of Environment*, 152, 217-234.

880 Zhu, Z., & Woodcock, C. E. (2014b). Continuous change detection and classification of land cover using
881 all available Landsat data. *Remote sensing of Environment*, 144, 152-171.

882 Zhu, Z., Woodcock, C. E., Holden, C., & Yang, Z. (2015b). Generating synthetic Landsat images based
883 on all available Landsat data: Predicting Landsat surface reflectance at any given time. *Remote*
884 *Sensing of Environment*, 162, 67-83.

885 Zhu, Z., Woodcock, C. E., Rogan, J., & Kellndorfer, J. (2012). Assessment of spectral, polarimetric,
886 temporal, and spatial dimensions for urban and peri-urban land cover classification using Landsat
887 and SAR data. *Remote Sensing of Environment*, 117, 72-82.