DEVELOPMENT OF A 2001 NATIONAL LANDCOVER DATABASE FOR THE UNITED STATES

Collin Homer, Chengquan Huang, Limin Yang, Bruce Wylie and Michael Coan
SAIC Corporation, USGS/EROS Data Center, Sioux Falls, SD 57198

Corresponding author: Collin Homer, 605-594-2714
homer@usgs.gov

ABSTRACT

Multi-Resolution Land Characterization 2001 (MRLC 2001) is a second-generation Federal consortium designed to create an updated pool of nation wide Landsat 5 and 7 imagery, and derive a second-generation National Land Cover Database (NLCD 2001). The objectives of this multi-layer, multi-source database are two fold: first to provide consistent land cover for all 50 States, and second to provide a data framework which allows flexibility in developing and applying each independent data component to a wide variety of other applications. Components in the database include the following: (1) normalized imagery for three time periods per path/row, (2) ancillary data, including a 30 m Digital Elevation Model (DEM) derived into slope, aspect and slope position, (3) per-pixel estimates of percent imperviousness and percent tree canopy, (4) 29 classes of land cover data derived from the imagery, ancillary data and derivatives, (5) classification rules, confidence estimates and metadata from the land cover classification. This database is now being developed using a mapping zone approach, with 66 zones in the continental United States and 23 zones in Alaska. Results from three initial mapping zones show single-pixel land cover accuracies ranging from 73-77%, imperviousness accuracies ranging from 83-91%, tree canopy accuracies ranging from 78-93% and an estimated 50% increase in mapping efficiency over previous methods. The database has now entered the production phase and is being created using extensive partnering in the Federal government with planned completion by 2006.

I. INTRODUCTION

Consistent, relevant land cover information at a national scale provides data for a wide variety of geographical analysis and applications. In the last decade, a major provider of land cover information within the Federal government has been the Multi-Resolution Land Characteristics Consortium (MRLC). The MRLC was originally formed in1993, to meet the needs of several Federal agencies (U.S. Geological Survey (USGS), Environmental Protection Agency (EPA), National Oceanic and Atmospheric Administration (NOAA), and U.S. Forest Service (USFS)) for Landsat 5 imagery and land cover information (Loveland and Shaw 1996). One of the products of this consortium was the completion of a successful mapping of the conterminous United States into the National Land Cover Dataset (NLCD 1992), derived from circa 1992 Landsat Thematic Mapper (TM) at the approximate Anderson et al., (1976) level II thematic detail (Vogelmann et al., 2001A). The continuing need for current Landsat based land cover data within the Federal government resulted in expanding the MRLC Consortium into a second stage effort called MRLC 2001 (more information at *www.mrlc.gov*). In addition to the USGS, EPA, NOAA and USFS, the MRLC 2001 Consortium also includes the Bureau of Land Management (BLM), National Aeronautics and Space Administration (NASA), National Park Service (NPS) and the Natural Resources Conservation Service (NRCS)[*].

The MRLC 2001 goals are twofold. First, a Landsat 7 and Landsat 5 image acquisition that includes multi temporal data processed to standard procedures for three dates per path/row (representing seasons) for the conterminous United States, Alaska, Hawaii and Puerto Rico. Second, a value-added database of land cover, called the National Land Cover Database 2001 (NLCD 2001), which is being generated across all 50 States and Puerto Rico using Landsat imagery and ancillary data.

The completion of the initial NLCD 1992 (Vogelmann et al., 2001A) created a TM pixel scale (30 m) data layer over the conterminous United States with approximately 9 billion pixels. During the 5 years of mapping required to complete this prototype product, many lessons were learned about quality of source data, objectivity of methods, and flexibility of products. This feedback, coupled with new MRLC 2001 member requirements, provided the guiding principles and research direction that culminated in the NLCD 2001 design. Principles included: (a) develop land cover products flexible enough for multiple users, (b) provide users with increased access to intermediate database products and derivatives, enabling local application, (c) develop methods that are as objective, consistent, and repeatable as possible, resulting in standardized land cover products that can be quickly updated, (d) constrain methods to those that are intuitive, simple, efficient, and transferable to others, and (e) ensure that the design of a second-generation land cover product maintains reasonable compatibility with NLCD 1992.

[*] The use of any trade, product or firm name is for descriptive purposes only and does not imply endorsement by the U.S. Government.

The NLCD 2001 foundation is a database approach to land cover (defined as multiple interlinked data layers that are useful either as individual components or in synergistic groupings) which builds upon past USGS database designs such as the global land cover database (Brown et al., 1999, Loveland et al., 2001), while providing the land cover data necessary to meet the vision of the *The National Map* (USGS 2001) currently being created by the USGS for the United States.

The NLCD 2001 also seeks to use this database approach to move beyond traditional remote sensing classification of land cover focused in specialized categories that meet only specific requirements. Historically, land cover products have often been developed according to specific project needs, with methods and results generally not designed to extrapolate to other areas or to crosswalk to different land cover schemes. These approaches have often resulted in remote sensing datasets and methods that develop categories that are difficult to compare (spatially and temporally) and have limited flexibility for other uses. This local product focus, historically often a limitation of technology and funding, has restricted the broad-scale development of remote sensing datasets, especially at nationwide scales. Product specific goals often result in potentially valuable intermediate data layers being discarded after the generation of the final product. These intermediate data layers (such as image transformations, ancillary information and classification rules) provide an untapped potential for flexible application if staged in an organized related database.

Continuing improvements in remote sensing data quality and availability, hardware capability and software algorithms have removed many of the technical barriers restricting the use of remote sensing data in more comprehensive and objective databases. We believe that NLCD 2001 offers an example of the incorporation of new technical improvements, balanced with product designs that offer flexibility in both production and use of the database. The result is a land cover database that is reasonably objective, consistent, and able to accommodate a variety of potential users and producers. We anticipate user access to this nationwide standardized database will foster additional exploration, development, application, and sharing of land cover information. This paper discusses the development, characteristics and implementation of this database.

II. DATABASE DEVELOPMENT

A science strategies team supported by the USGS and EPA did the development of NLCD 2001 at the USGS EROS Data Center (EDC) beginning in 1999. Four study sites representing different types of land cover in the conterminous United States were selected, and were the focus of research trials involving various classification methods (Figure 1). Two sequential Landsat 7 path/rows were selected to represent these sites, which included Virginia (eastern deciduous forest and agriculture), Nebraska/South Dakota (midwest crop/prairie/pasture), Utah (Rocky Mountain and Great Basin shrubs/forests and irrigated agriculture) and Oregon (coastal forests, agriculture and shrublands). Methods developed in research trials at these study sites were assumed to extrapolate to the conterminous United States, and possibly to Alaska, Hawaii, and Puerto Rico. Following 3 years of comprehensive review and research by this team, the database design for NLCD 2001 represents the efforts to follow the guiding principles outlined by MRLC 2001.

III. DATABASE CHARACTERISTICS

The NLCD 2001 database is presented in Figure 2. Stratified by mapping zones, the database consists of the following components: (1) normalized tasseled cap (TC) transformations of Landsat 7 imagery for three time periods per path/row (early, peak and late growing seasons) plus the thermal band calibrated to temperature, (2) ancillary data layers including a 30m Digital Elevation Model (DEM) and derivatives of slope, aspect and slope position, (3) per-pixel estimates of percent imperviousness and percent tree canopy, (4) 29 classes of land cover data derived from the imagery, ancillary data and derivatives, (5) classification rules, classification confidence and metadata describing the land cover classification. The rest of the paper will focus on describing the characteristics of each component and report the results of initial classifications.

Mapping Zone Delineation

Originally, NLCD 1992 was mapped in zones determined by U.S. Federal region boundaries. These were unrelated to the biogeography of land cover and caused difficulties in mapping because mosaic boundaries included widely disparate land cover types. This experience led to a focus on an improved regional stratification method for NLCD 2001 as a means to stage both the components and the derived products of the database. Because mapping over large landscapes typically involves many satellite scenes, multi scene mosaicking has often been used to group scenes into a single file for classification. This approach can potentially optimize both classification and edge matching (Homer et al., 1997).

However, large multi scene mosaics create a variety of spectral gradients within the file, and these files are subsequently useful only as a mosaicked unit. Spectral gradients typically represent patterns of physiographic, phenologic, solar, atmospheric and instrument influences within and between remotely sensed images. The degree to which this variability can be isolated in local context largely determines the success of the land cover classification. A common method of isolating spectral gradients is to stratify landscapes into sub regions of similar biophysical and spectral characteristics. This process is not new to remote sensing and has been widely used as a method to improve accuracy (White et al., 1995; Lillisand 1996). For example, Bauer et al., (1994) showed that overall classification accuracy could potentially be improved by 10 to 15 percent using physiographic regions for stratification.

The underlying concept of mapping zone delineation is a pre-classification division of the landscape into a finite number of units that represent relative homogeneity with respect to landform, soil, vegetation, spectral reflectance and image footprints at a project scale that is affordable. Five general concepts are useful in defining mapping zones; economics of size, type of physiography, potential land cover distribution, potential spectral uniformity, and edge-matching issues (Homer and Gallant 2001). We assume that application of mapping zones as a pre-classification stratification method for NLCD

2001, will maximize spectral differentiation, provide a means to facilitate partitioning the workload into logical units, simplify post classification modeling, improve classification accuracy and minimize edge matching.

The development of mapping zones across the conterminous United States included an initial review of project scope, which determined that approximately 60-70 zones would be the appropriate grain size for staging NLCD 2001. Initial mapping zone boundaries were based on 83 level III ecoregions developed by Omernik (1987). These initial boundaries were displayed over two principal data layers, NLCD 1992 and Advanced Very High Resolution Radiometer (AVHRR) normalized greenness maps for modification. These data layers provided a landscape overview of interpreted land cover and gross vegetation phenology patterns and provided the context to further refine the initial Omernik boundaries on 1:5,000,000 scale paper maps. Paper map boundaries were subsequently crafted into a digital file by onscreen digitizing with NLCD 1992 as the background. Initial digital boundaries are refined over full-resolution TM data as each zone is actually mapped to create local line interpretation relevant at the single-pixel scale. Mapping zones were developed for both the conterminous United States and Alaska (Figure 3).

Database Imagery

-Scene Selection

The strategy developed for nationwide Landsat imagery selection was designed to meet the requirements of three acquisition dates for each Landsat path/row covering early, peak and late vegetation green-up (Yang et al., 2001A). Scene selection criteria were established using multi temporal greenness as an indicator of vegetation phenology. Information on vegetation phenology was derived from the multi-temporal normalized difference vegetation index (NDVI) data of the conterminous United States acquired by the AVHRR from 1994 to 1998 (Swets et al., 1999). Landsat potential date selection “windows” were identified using the average NDVI annual trajectories, qualified by proportions of land cover types in each path/row. This method provided a general guide of optimal “windows” for selection to maintain regional consistency. Exceptions to acquiring images outside the date windows were granted only when good-quality cloud-free data were not available. Overall, this strategy has been successful, providing a reasonably objective framework to populate the nationwide image database.

It was initially assumed that Landsat 7 ETM+ would be the exclusive data source for the image database. However, the addition of Landsat 5 TM to Federal Government control, with its additional pool of cloud-free imagery, created a unique opportunity to populate the database with additional selections to better represent ideal image acquisition windows. Special processing to ensure the compatibility of Landsat 5 to Landsat 7 ETM+ data is explained in the next section. Currently, Landsat 5 TM imagery comprises only 14% of the database.

-Preprocessing

Challenges to large-scale, multi frame, satellite-based land cover characterization include consistent geometric correction, normalizing noise arising from atmospheric effect, adjusting for changing illumination geometry, and minimizing instrument errors inherent when using multiple frames of imagery. Such geometric and radiometric error can hinder the ability to derive land surface information reliably and consistently.

For MRLC 2001, images are geometrically corrected using cubic convolution resampling in a single step from Level 0 data to Level 1GT, which provides terrain correction. Terrain correction is performed using the USGS 1-arc second National Elevation Dataset (NED) (Gesch et al., 2002) to improve geo location accuracy. The selection of cubic convolution as a resampling strategy was based largely on the superior spatial accuracy it provides over nearest neighbor resampling (Shilen 1979, Park and Schowengerdt 1982). This is of special concern when stacking multiple dates across many path/rows, as is the case with NLCD 2001. For Landsat 7 ETM+ the visible and infrared bands (bands 1-4, 5,7) are resampled to a 30 m spatial resolution; the panchromatic band (band 8) is resampled to 15 meters and the thermal band (band 6) to 60 meters. For Landsat 5, the visible and infrared bands (bands 1-4, 5,7) are resampled to a 30-m spatial resolution, and the thermal band (band 6) to 90-m resolution.

Great efforts have been made to minimize radiometric noises due to instrument errors for standard image products of Landsat 7 (Irish 2001). Noise due to the influence of the atmospheric and illumination geometry can be normalized using several approaches. For MRLC 2001, Landsat 7 images are first radiometrically corrected using standard methods at the USGS EDC to eliminate band bias and gain anomalies (Irish 2001). For Landsat 5, a radiometric conversion to Landsat 7 is first performed using the inverse of coefficients developed by Vogelmann et al., (2001B) for Landsat 7 to Landsat 5 conversions. Initial tests on NLCD 2001 sites indicated this provided an adequate radiometric calibration of Landsat 5 data (error rates usually around 2-3%), enabling the mixing of both Landsat 7 and Landsat 5 data in a single mosaic.

Next, Landsat images are converted to at-satellite reflectance for the six reflective bands (not the panchromatic) and to at-satellite temperature for the thermal band according to Markham and Barker (1986) and the Landsat 7 Science Data Users Handbook (Irish 2001). Considering the tremendous volume of imagery being processed (1,780 path/rows) and the relative uncertainty of algorithms currently available, atmospheric and topographic normalizations are not performed because of their potential to introduce confounding error. Only first order normalization conversion to at-satellite reflectance is done on clear and near cloud-free images. This conversion algorithm is physically based, automated, and does not introduce significant errors to the data (Huang et al., 2002A). Initial tests have shown that this method, which normalizes multi scene noise, coupled with the intelligent scene selection strategy, provides a reasonable preprocessing method for such a large database. In many areas this method will allow assembling of multi-scene datasets without using traditional histogram-matching mosaicking (Figure 4).

Mapping zone image mosaics are currently produced using only first-order normalized imagery with no histogram matching or adjustment. Although this method provides an important first step in standardizing imagery, some atmospheric, phenological and topographic noise still remains among images. However, more importantly the lineage to the original scenes from the database are preserved in the mosaic.

-Spectral Data Transformation

Potential use of portions of a nationwide, three-date, Landsat TM database would require enormous hardware storage capability for a user. Possibilities were explored for optimal ways to distill original resolution TM bands into spectral-efficient transformations without losing important information. Principle Component Analysis (PCA) derivatives were assumed to be the most efficient transformation for compressing spectral information. However, PCA was not considered a viable method for image compression because of its interpretation difficulty, especially when comparing image to image. Tests and trials using indices such as NDVI, Soil Adjusted Vegetation Index (SAVI), Leaf Area Index (LAI) and Tasseled Cap (TC) transformations were compared against PCA results. A universal PCA transformation was derived from random pixels from multiple dates and path/rows. The percent of the total spectral and thermal variance explained by each principal component was multiplied by the percent of the variance explained by each spectral index (R_2 from linear regression) to quantify the percent of spectral variance explained by each tested index. Tests showed that TC offered the best potential surrogate to PCA retaining 98% of potential PCA all-band spectral variance information. More importantly, TC offers the additional advantage of providing standardized output layers of brightness, greenness and wetness that are linked to scene physical characteristics and comparable from image with image.

This new TC transformation is applicable to Landsat 7 at-satellite reflectance normalized scenes described above was developed from 10 ETM+ scenes representing a variety of landscapes across the United States in both leaf-on and leaf-off seasons (Huang et al., 2002A). The brightness, greenness and wetness of the derived transformation collectively explained more than 97% of the spectral variance of individual scenes used in this study.

Database Ancillary Information

Successful land cover mapping typically needs ancillary data for improvement. The type of ancillary data available and the method used to classify them both play a large role in the success of the classification. For NLCD 2001, the use of decision and regression tree algorithms for classification of the database allows ancillary data full weighting in the classification process. This highlights the need for consistent and meaningful ancillary data sources. Ancillary data layers that have been standardized for use in the database include both the NED (Gesch et al., 2002) DEM and three DEM derivatives including slope, aspect and a positional index. Slope is calculated in degrees, aspect is calculated into 16 directional classes and the slope positional index is based on a 7x7-weighted filter modified from Dikau et al., (1995). Additional ancillary data, such as population density data, buffered roads, NLCD 1992 and NOAA City Lights, are used for urban masking (Yang et al., 2002). Other data, such as the National Wetland Inventory or other regionally available data, may be carefully applied in appropriate mapping zones if national consistency can be maintained.

Database Derivatives

-Imperviousness

Impervious surfaces refers to impenetrable surfaces such as rooftops, roads or parking lots. Quantification of imperviousness can offer a relatively objective measure of urban density and provide a forum for its classification**.** For NLCD 2001, imperviousness was chosen as the surrogate for the urban intensity classification in an effort to improve the precision of urban characterization used in the original NLCD 1992.

Modeling empirical relationships between imperviousness and Landsat data is accomplished using regression tree techniques. Several one-meter digital orthophoto quadrangles are used for each Landsat scene to derive reference impervious data needed for calibrating the relationships between percent imperviousness and Landsat spectral data, which are then modeled using a commercial regression tree algorithm called Cubist. The models are then applied to all pixels in a mapping zone to produce a per-pixel estimate of imperviousness in urban areas (Yang et al., 2002). This procedure quantifies the spatial distribution of impervious surfaces as a continuous variable for urban areas from 1 to 100%, and offers a consistent and repeatable method to characterize urban areas across the Nation. This data layer is then masked to ensure only urban pixels are included and thresholded (Table 1) into NLCD 2001 urban classes and inserted into the land cover. Imperviousness information will be available as an independent product of NLCD 2001.

-Tree Canopy

Forest canopy density is of great interest to a variety of scientific and land management users. The original NLCD 1992 classification provided four forest categories but made no distinction in forest canopy density. For NLCD 2001, a strategy for estimating tree canopy density at a spatial resolution of 30 m was developed (Huang et al., 2001). This strategy is similar to the method used to derive imperviousness, and is based on empirical relationships between tree canopy density and Landsat data, established using regression tree techniques. Several one-meter digital orthophoto quadrangles for each Landsat path/row are required to derive reference tree canopy density data needed for calibrating the relationships between canopy density and Landsat spectral data. As with the imperviousness data layer, the regression tree algorithm Cubist is used to develop the models and output a per-pixel estimate of tree canopy for all pixels. To aid the utility of the canopy estimate as an independent data layer, a non-forest mask is created to mask obvious non-forest pixels from the prediction. This procedure quantifies spatial distribution of tree canopy as a continuous variable from 1 to100%, and will be available as an independent product of NLCD 2001.

-Land Cover

There are numerous algorithms for classifying satellite images. Potential methods reviewed for NLCD 2001 included spectral clustering, expert systems, neural networks and decision tree classifiers. NLCD 1992 classification was based on a several-step method of unsupervised clustering, using both pre-classification and post-classification stratification with ancillary data, and manual editing to complete the work (Vogelmann et al., 2001A). For NLCD 2001, a method that optimally classifies many database layers in a single step, with the ability to document this relationship in a rule base was highly desirable. Decision tree classification (Breiman et al., 1984, Lawrence and Wright 2001) was the method chosen for NLCD 2001. Advantages it offers include: (1) it is non-parametric and therefore independent of the distribution of class signature, (2) it can handle both continuous and nominal data, (3) it generates interpretable classification rules, and (4) it is fast to train and often as accurate as, or even slightly more accurate than many other classifiers. The commercial decision tree program used in this case study, C5, employs an information gain ratio method in tree development and pruning (Quinlan 1993), and has many advanced features including boosting and cross-validation.

For NLCD 2001, decision tree classification offers an efficient, robust method to classify large quantities of information in documentable form. Additionally, decision trees allow export of mutually exclusive rules generated by the classification into generic textual rule sets allowing users access to classification parameters. They also allow the generation of a classification confidence map, which as part of NLCD 2001 metadata, allow users more feedback on the reliability of the land cover information. Additionally, trees enable the output of a "node" map, which spatially shows where pixels for each decision tree node are located (similar to spectral clusters). Landcover users thus gain metadata feedback on what input layers generated the land cover classification, a spatial map of which tree nodes made the prediction, and a spatial confidence map showing how confident the classifier was in making that prediction. This comprehensive metadata approach will allow users access to classification reasoning and will potentially allow local modification of the classification database for more specific applications.

One of the challenges in land cover mapping using a supervised method over large areas is the need for adequate reference data. Decision trees are a supervised method of classification and require extensive well-balanced training data both spatially and categorically to perform adequately. Scarcity of reliable reference data and subsequent lack of consistency often limit the accuracy of land cover information derived from satellite imagery. Reference data for NLCD 92 were collected using a combination of aerial photographs and fieldwork. For NLCD 2001, additional nation-wide training data consistency was sought by partnering with other Federal programs.

One example has been the successful collaboration with the Forest Inventory Analysis (FIA) program of the USFS. The FIA has a mandate to collect and report information on status and trends in the Nation's forested resources. FIA plots represent a probability based sampling of the Nation's land, and detailed information on forest status and structure is collected periodically at each plot through intensive field work. Tests have shown that with minimal effort, this dataset can be reorganized for use as training data for NLCD 2001 (Huang et al., 2002B). The plot data collected through the FIA program provides a high quality reference dataset for the NLCD 2001, allowing a more consistent forest classification nationwide. In turn, the FIA then benefits from NLCD 2001 forest classes providing more optimal initial stratification for statistical designs.

Overall, training data will be collected from existing federal and state programs, complemented with new field collection in areas where current data are not available.

For NLCD 2001, 29 classes of land cover are targeted for mapping (Table 1), with 13 new classes from NLCD 1992. Four of these new classes are unique to Alaska and nine classes are unique to coastal zones. For the continental United States, water, forest, shrub, herbaceous and wetland classes are nearly identical to NLCD 1992 definitions, with agriculture, urban and barren classes having slightly more adjustment.

Database Metadata

In order to provide comprehensive documentation and to enable further use of the land cover database, two additional raster layers and one text file are generated to provide users feedback on data quality and classification lineage. The first raster layer is a confidence map generated by the decision tree algorithm that reports a per-pixel confidence in the classification based on the training data. The second raster layer is a tree "node" map that allows users to spatially observe which pixels in the output are represented by which nodes in the decision tree (this output appears similar to an unsupervised cluster map). The text file is intended to be used as a companion to the node map, and contains logical statements by node that document classification parameters for each input layer used by the decision tree. This text file can be imported into many geospatial software packages and used as classification rules. These files enable users to potentially scrutinize and modify regional pixel areas. It is important to note that because decision tree boosting is used in the initial creation of the land cover product, metadata files are generated in a second step using an un-boosted decision tree trained on the initial product. Hence, the metadata only provide a very close approximation of the original single pixel land cover product.

Database Validation

NLCD 92 was validated using aerial photographs within a sampling design incorporating three levels of stratification and a two-stage cluster sampling protocol (Stehman and Czaplewski 1998, Zhu et al., 1999,Yang et al., 2001B). This process produced a credible assessment for users, but also created a significant time lag between production and assessment, thus limiting critical feedback. Both decision and regression trees offer cross validation as an option to initially assess product quality. Cross-validation can provide relatively realistic accuracy estimates when using reference data samples that are statistically valid for both training and accuracy assessment (Michie et al., 1994). For an *N*-fold cross-validation, the training data set is divided into *N* subsets. Accuracy estimates are derived by using each subset to evaluate the predictions developed using the remaining training samples, and their average value represents the results developed using all reference samples. For NLCD 2001, initial tests revealed comparable accuracies between independent data assessment and cross-validated estimates (Huang et al., 2003). Cross validation for land cover will be used to determine the accuracy of preliminary results, with an independent assessment deployed later. Regression tree cross validation will provide the main assessment tool for accuracy estimates of tree canopy and imperviousness modeling.

IV. IMPLEMENTATION

The full database described above has been completed in zones 16, 41 and 60 (Figure 5). Specifics of the development are described below by product.

Imperviousness and Tree Canopy Classification

Methods for imperviousness described by Yang et al., (2002) and tree canopy methods described by Huang et al., (2001), were applied in two mapping zones by Earth Satellite Corporation (EarthSat) through Greenhorne & O'Mara Inc. under USGS contract number 010112C0012. The USFS Remote Sensing Mapping Applications Center (RSAC) in Salt Lake City and the EDC applied identical methods in zone 41. For imperviousness in zone 60, 20 DOQ subsets were used to generate training data, with 16 subsets in zone 16, and 11 subsets in zone 41. Forest canopy required 16 DOQ subsets of training data in zone 16, 17 subsets in zone 41, and 21 subsets in zone 60. Both imperviousness and canopy estimates at the 30-m resolution were developed using the Cubist regression tree algorithm. Typical input prediction data layers included three seasonal TC Landsat 7 images or spectral bands, the leaf-on thermal band, and in some cases the elevation, aspect and slope. Per-pixel imperviousness and tree canopy estimates for each zone were generated and assessed using cross-validation (Table 2). Canopy results reveal mean absolute errors (mean of the absolute difference between predicted and actual values) from 8.4% to 14.1%, with correlation coefficients (r) between predicted and actual values ranging from 0.78 to 0.93 (Figure 6). Imperviousness results reveal mean absolute errors from 4.6% to 7%, with r-values from 0.83 to 0.91 (Figure 7).

Land cover classification

Land cover was derived from a combination of image and ancillary layers using the C5 decision tree program. Reference data for zones were collected from combined sources. The majority of forested reference data were provided for each region through a unique pilot agreement with the USFS FIA. Incorporating this evenly sampled dataset improved forest mapping considerably, and provided reliable cross-validation estimates. Other reference data sources in zone 16 included EDC collected data, USFS Fire Science Lab of the Rocky Mountain Research Station, and the Utah GAP Analysis program of Utah State University. In zone 60, field data were also contributed by the State of Delaware and EDC. In zone 41, a unique agreement with the USDA National Resource Inventory program (NRI) enabled access to their resource inventory data for the entire zone. NRI is a nationwide assessment program similar to FIA, with the mandate to focus on agriculture and wetland areas.

Input prediction data included 26 layers of multi-temporal spectral and ancillary data in zone 60, 20 layers in zone 16, and 16 layers in zone 41. A total of 12 classes were mapped in zone 60 (Figure 8) using a hierarchical approach that mapped forest classes separately from agriculture and wetland. A total of 18 classes in zone 16 and 14 classes in zone 41 were mapped also using the same hierarchical approach. Cross-validation accuracies (Table 3) for the three zones ranged from 72.6 in zone 41 to 77.2 in zone 60, with standard errors ranging from 1.2 % in zones 16 and 60 to 2.1% in zone 41. Several iterations of decision trees are typically required for each zone to finalize the land cover.

Database Partners

A direct result of the utility and flexibility of the NLCD 2001 database has been the further development of extensive partnerships with Federal and State agencies, representing a good example of how government agencies can work together to achieve complimentary objectives (Figure 9). For example, NLCD 2001 has provided a way to further combine mapping efforts within the USGS by synergistic mapping with the GAP Analysis program, as well as combining mapping efforts with other agencies such as NOAA's (Coastal Change and Analysis) CCAP program (see Table 1).

USGS, EDC serves as the primary catalyst to manage database development, maintain the quality and consistency of database products, preprocess data ingredients, provide training on classification methods, supervise data generation and quality assurance and provide dissemination. Other federal partners provide direct support in generating land cover, imperviousness, and tree canopy classifications, which are then incorporated into NLCD 2001.

V. CONCLUSION

Based on land cover results from the three zones described in this paper, it was estimated the new NLCD 2001 method produced about a 50% gain in mapping efficiency with comparable or improved accuracies over NLCD 1992 methods. Additionally, both imperviousness and canopy data provided value independent of the land cover.

Based on these initial results we believe the NLCD 2001 database can provide a comprehensive set of data layers with the potential to foster further exploration, development, application and sharing of land cover information by users at national and regional scales. The standardized nature of each data component can allow users the ability to develop data applications that use layers either synergistically or individually. For example, imperviousness can potentially be used not only as a way to classify developed land, but also in water run-off models, green space calculations and urban planning scenarios. Tree canopy can be intersected with NLCD 2001 forest classes to provide canopy categories by density. Further, the consistency of these data layers will allow direct comparison from place to place across the Nation, increasing the utility of potential applications.

The database framework also can provide users flexible access and interaction with the individual data components and land cover products. Spatial and textual metadata generated from land cover product development will allow users the ability to download both database ingredients and metadata for potential local evaluation. Conceptually, a potential user could modify land cover model parameters directly by manipulating rule-set parameters according to more local information. In this scenario, NLCD 2001 acts as a framework to provide standardized ingredients and a general "recipe" empowering less sophisticated users to generate local value-added land cover without extensive preparation. Further, this database could provide a common "language" for users to access, compare, and model intermediate remote sensing information for the U.S., thus capturing the full potential of the database model.

The production of NLCD 2001 will be implemented in a phased approach using the mapping regions developed by the USGS. Full production is now in development, and contingent on funding from MRLC 2001 partners. Completion is targeted for 2006. MRLC 2001 will welcome additional cooperation from Federal, State and other partners.

AKNOWLEDGEMENTS

This study is made possible in part by the SAIC Corporation under U.S. Geological Survey contract 1434-CR-97-CN-40274. This research reflects an enormous team effort with important contributions made by many individuals including (in alphabetical order) Alisa Gallant, Kent Hegge, Tom Loveland, Russell Rykhus, Nick Van Driel, James Vogelmann and Jane Zhang. James Wickham of the EPA and James Vogelmann authored an initial MRLC 2001 white paper that provided early guidance. Space Imaging's Portland office conducted key research on potential shape and texture methods (under USGS contract number 010112C0012). Earthsat Corporation developed a working version of canopy and imperviousness (under USGS contract number 010112C0012) through Greenhorne & O'Mara Inc at the mapping zone scale. Cooperation with the USFS FIA under the direction of Dr. Andy Gillespie was key in developing forest classifications in the reported zones. Cooperation with USDA NRI, under the direction of Dr. Jeff Goebel, and Dave Heinzen of Minnesota DNR was key in developing the classification for zone 41. We also wish to express our gratitude to the NASA Science Data Purchase Program for providing the Ikonos data used in the study.

REFERENCES

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

Bauer, M.E., T.E. Burk, A.R. Ek, P.R. Coppin, S.D. Lime, T.A. Walsh, D.K. Walters, W. Befort, and D.F. Heinzen, 1994. Satellite inventory of Minnesota forest resources. *Photogrammetric Engineering and Remote Sensing* 60: 287-298.

Breiman, L., Friedman, J.H., Olshend, R.A., and Stone, C.J., 1984. *Classification and regression trees*, Belmont, California, Wadsworth International Group, 358 p.

Brown, J., T. Loveland, D. Ohlen and Z. Zhu, 1999. The global land-cover characteristics data-base: The users' perspective. *Photogrammetric Engineering and Remote Sensing* 65:1069-1074.

Dikau, R., Brabb, E. E., Mark, R. K., and Pike, R. J. (1995): Morphometric landform analysis of New Mexico. In: *Advances in geomorphometry - Proceedings of the Walter F. Wood Memorial Symposium*, volume 101 of *Zeitschrift für Geomorphologie,* Supplement Band: 109-126.

Gesch, D., M. Oimoen, S. Greenlee, C. Nelson, M. Steuck and D. Tyler, 2002. The National Elevational Dataset, *Photogrammetric Engineering and Remote Sensing* 68:5-11.

Homer, C.G., R.D. Ramsey, T.C. Edwards, Jr., and A. Falconer, 1997. Land cover-type modeling using a multi-scene Thematic Mapper mosaic, *Photogrammetric Engineering and Remote Sensing* 63:59-67.

Homer, C.G. and A. Gallant. 2001. Partitioning the conterminous United States into mapping zones for Landsat TM land cover mapping, USGS Draft White Paper, available at http://landcover.usgs.gov.

Huang, C., L. Yang, B. Wylie, and C. Homer, 2001. A strategy for estimating tree canopy density using Landsat 7 ETM+ and high resolution images over large areas, *Third International Conference on Geospatial Information in Agriculture and Forestry*; November 5-7, 2001; Denver, Colorado. CD-ROM, One disk.

Huang, C., B. Wylie, C. Homer, L. Yang, and G. Zylstra, 2002A. Derivation of a Tasseled cap transformation based on Landsat 7 at-satellite reflectance, *International Journal of Remote Sensing*, Vol. 23:No. 8, 1741-1748.

Huang, C., L. Yang, C. Homer, M. Coan, R. Rykhus, Z. Zhang, B. Wylie, K. Hegge, Z. Zhu, A. Lister, M. Hoppus, R. Tymcio, L. DeBlander, W. Cooke, R. McRoberts, D. Wendt, and D. Weyermann, 2002B. Synergistic use of FIA plot data and landsat 7 etm+ images for large area forest mapping, *The Thirty-fifth Annual Midwest Forest Mensurationists and the Third Annual FIA Symposium*; October 17 - 19, 2001; Traverse City, MI.

Huang, C., Homer, C., and Yang, L., 2003. Regional forest land cover characterization using Landsat type data, *Methods and Applications for Remote Sensing of Forests: Concepts and Case Studies* (M. Wulder and S. Franklin eds.), Kluwer Academic Publishers, Boston, p. 389-410.

Irish, R.R., 2001. Landsat 7 science data user's handbook, Report 430-15-01-003-0, National Aeronautics and Space Administration, URL:http://ltpwww.gsfc.nasa.gov/IAS/handbook/handbook_toc.html.

Lawrence, R.L. and A. Wright, 2001. Rule-based classification systems using classification and regression tree (CART) analysis, *Photogrammetric Engineering and Remote Sensing* 67: 1137-1142.

Lilliesand, T.M. 1996. A protocol for satellite-based land cover classification in the upper Midwest, *Gap Analysis: A Landscape Approach to Biodiversity Planning*, (Michael Scott, Timothy H. Tear and Frank W. Davis eds.). Proceedings of the ASPRS/GAP Symposium, Charlotte, NC , 320 pp.

Loveland, T.R., B.C. Reed, J.F. Brown, D.O. Ohlen, Z. Zhu, L. Yang, and J.W. Merchant, 2001. Development of a global land cover characteristics database and IGBP DISCover from 1-km AVHRR data, *International Journal of Remote Sensing,* 21(6/7):1,303-1,330.

Loveland, T.R. and D.M. Shaw, 1996. Multiresolution land characterization: building collaborative partnerships, *Gap Analysis: A Landscape Approach to Biodiversity Planning* (J.M. Scott, T.Tear, and F. Davis eds.), Proceedings of the ASPRS/GAP Symposium, Charlotte, NC, pp. 83-89.

Markham, B.L. and J.L. Barker, 1986. Landsat MSS and TM post-calibration dynamic ranges, exoatmospheric reflectances and at-satellite temperatures, *EOSAT Landsat Technical Notes*, 1:3-8.

Michie, D., D.J. Spiegelhalter, and C.C. Taylor, (Eds.), 1994. *Machine learning, neural and statistical classification,* New York: Ellis Horwood, 289 pp.

Omernik, J.M., 1987. Ecoregions of the Conterminous United States, Map (scale 1:7,500,000), *Annals of the Association of American Geographers* 77(1): 118-125.

Park, Stephen K. and Robert A. Schowengerdt, 1982. Image reconstruction by parametric cubic convolution, *Computer Vision, Graphics and Image Processing* 23:258-272.

Quinlan, J. R., 1993. *C4.5 programs for machine learnin,* San Mateo, California: Morgan Kaufmann Publishers.

Shilen, S., 1979. Geometric correction, registration and resampling of Landsat imagery. *Canadian Journal of Remote Sensing* 5(1):75-89.

Stehman, S.V. and R.L. Czaplewski, 1998. Design and analysis for thematic map accuracy assessment: Fundamental principles, *Remote Sensing of Environment* 64:331-344.

Swets, D.L., B.C. Reed, J.R. Rowland, S.E. Marko, 1999. A weighted least-squares approach to temporal smoothing of NDVI, *Proceedings of the ASPRS 1999 Annual Convention*, Portland, Oregon. American Society for Photogrammetry and Remote Sensing, Bethesda, Maryland, CD-ROM, One disc.

U.S. Geological Survey, 2001, *The National Map: Topographic Mapping for the 21st Century*: Reston, Va., Office of the Associate Director of Geography, U.S. Geological Survey.

Vogelmann, J.E., S.M. Howard, L. Yang, C.R. Larson, B.K. Wylie, and J.N. Van Driel, 2001A. Completion of the 1990's National Land Cover Data Set for the conterminous United States, *Photogrammetric Engineering and Remote Sensing* 67:650-662.

Vogelmann, J.E., D. Helder, R. Morfitt, M.J. Choate, J.W. Merchant and H. Bulley, 2001B. Effects of Landsat 5 Thematic Mapper and Landsat 7 Enhanced Thematic Mapper Plus radiometric and geometric calibrations and corrections on landscape characterization, *Remote Sensing of Environment* 78:55-70.

White, J.D., G.C. Kroh, C. Glenn, and J.E. Pinder 3rd., 1995. Forest mapping at Lassen Volcanic National Park, California, using Landsat TM data and a geographical information system, *Photogrammetric Engineering and Remote Sensing* 61: 299-305.

Yang, L., C. G. Homer, K. Hegge, C. Huang, B. Wylie, and B. Reed, 2001A. A Landsat 7 scene selection strategy for a National Land Cover Database, *Proceedings of the IEEE 2001 International Geoscience and Remote Sensing Symposium*, Sydney, Australia, CD-ROM, One Disc.

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

Yang, L, C. Huang, C. Homer, B. Wylie, and M. Coan., 2002. An approach for mapping large-area impervious surfaces: Synergistic use of Landsat 7 ETM+ and high spatial resolution imagery, *Canadian Journal of Remote Sensing* 29(2):230-240.

Zhu, Z. L. Yang, S.V. Stehman, and R.L. Czaplewski, 1999. Accuracy assessment for the U.S. Geological Survey regional land cover mapping program: New York and New Jersey Region, *Photogrammetric Engineering and Remote Sensing*, 66:1425-1435.

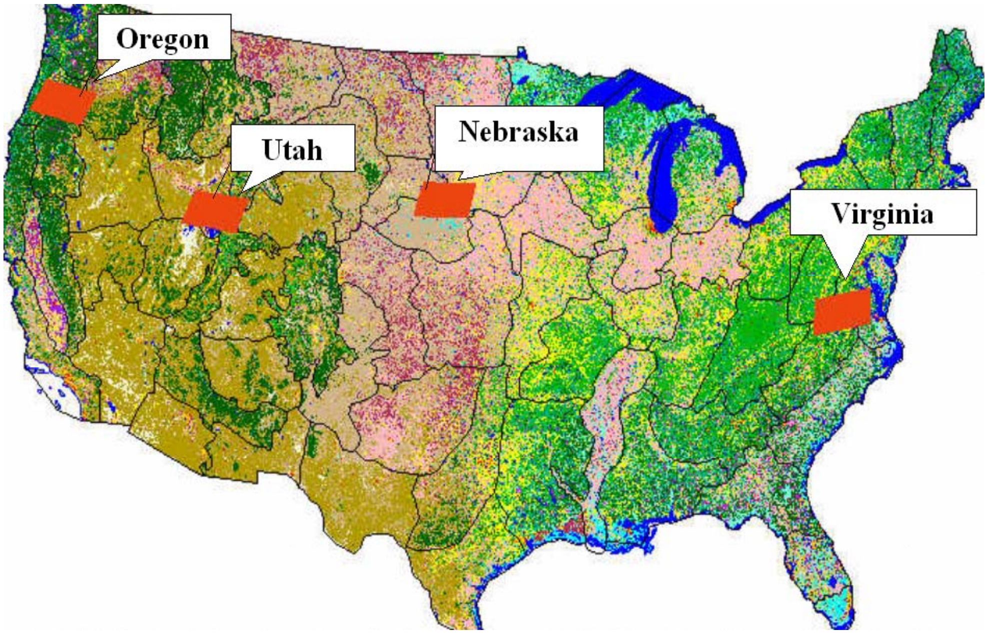
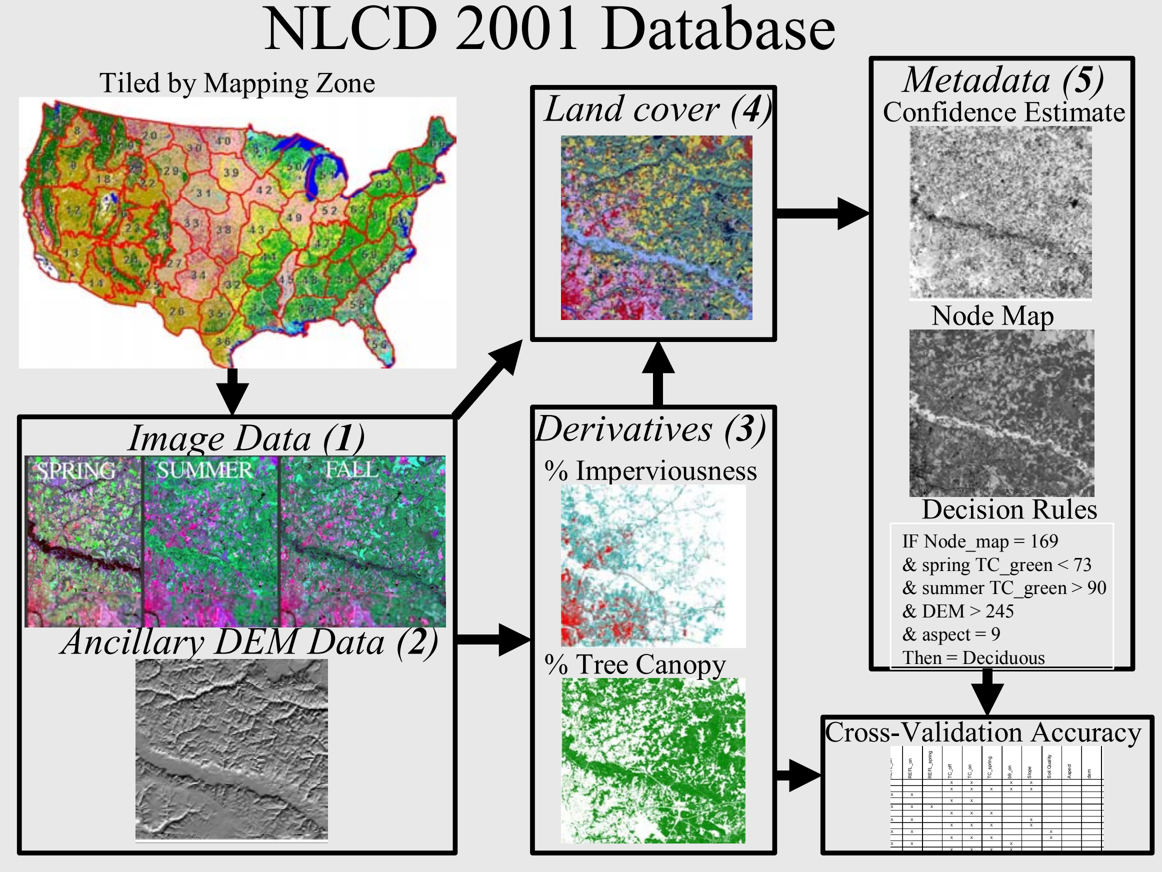


Figure 1. Four mapping strategy study sites used to develop nation-wide methods



NLCD 2001 Database

Tiled by Mapping Zone

Image Data (1)

SPRING SUMMER FALL

Ancillary DEM Data (2)

Derivatives (3)

% Imperviousness

% Tree Canopy

Land cover (4)

Metadata (5)

Confidence Estimate

Node Map

Decision Rules

IF Node_map = 169
& spring TC_green < 73
& summer TC_green > 90
& DEM > 245
& aspect = 9
Then = Deciduous

Cross-Validation Accuracy

REFL_off	REFL_on	REFL_spring	TC_off	TC_on	TC_spring	b9_on	Slope	Soil Quality	Aspect	dem
			x	x		x	x			
			x	x	x	x	x			
x	x									
			x	x						
x	x	x								
			x	x	x					
x	x						x			
			x	x	x		x			
x	x							x		
			x	x	x			x		
x	x					x				
			x	x	x	x				

Figure 2. The NLCD 2001 Database model, displaying both the processing flow and the characteristics of major data components.

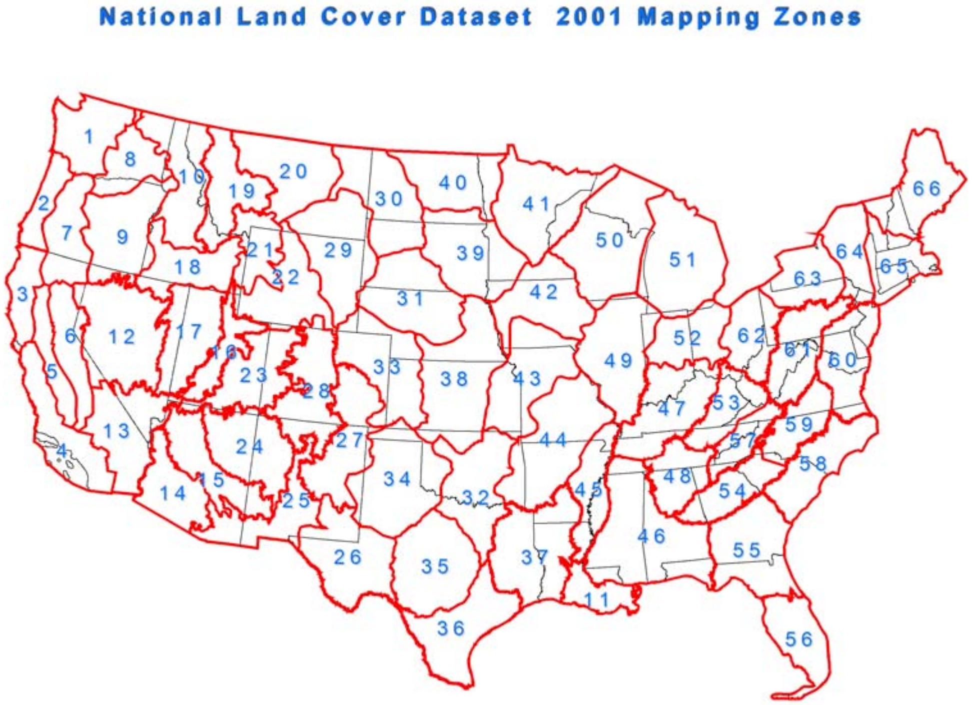
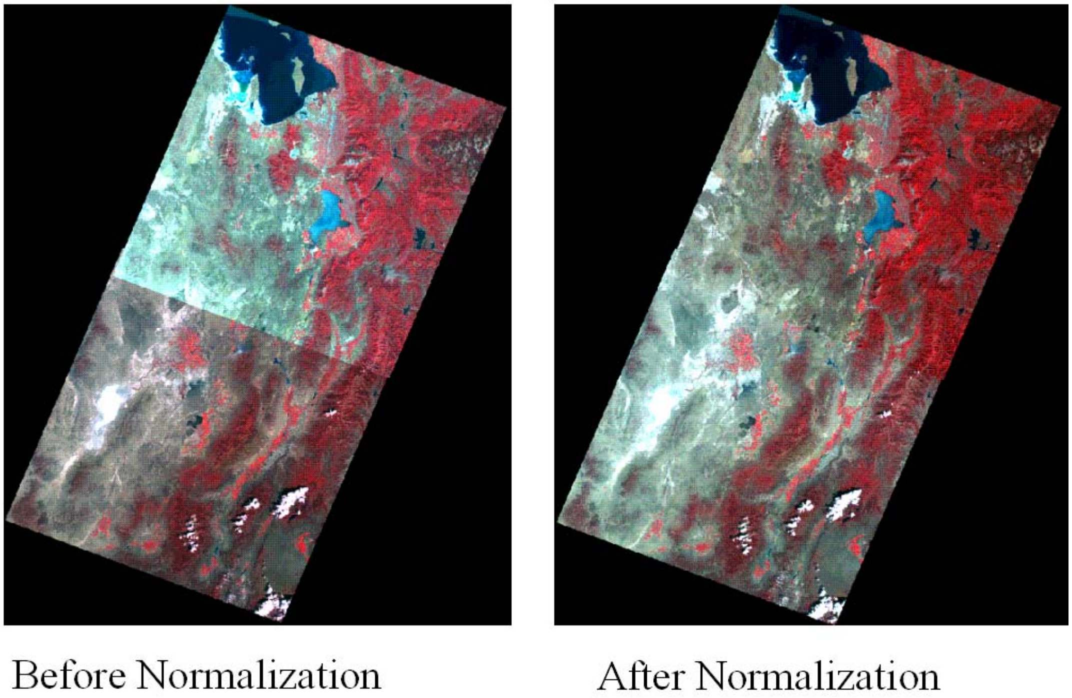


Figure 3. NLCD 2001 Mapping Zones displayed over State boundaries.



Before Normalization

After Normalization

Figure 4. Example of image normalization from “Top of Atmosphere” correction.

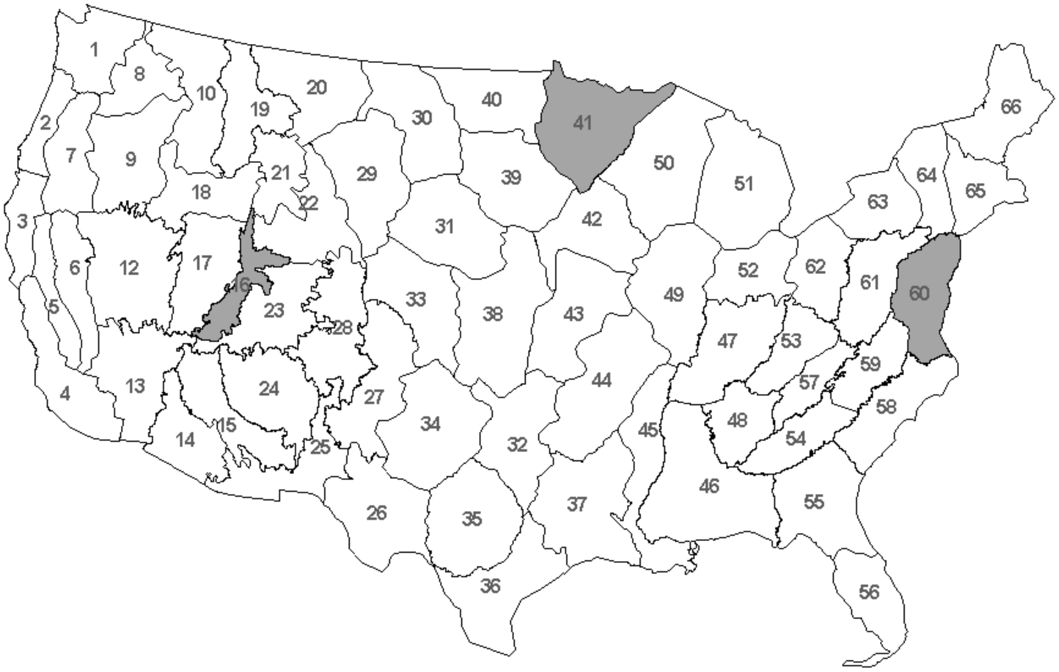


Figure 5. NLCD 2001 Mapping Zones with the completed database.

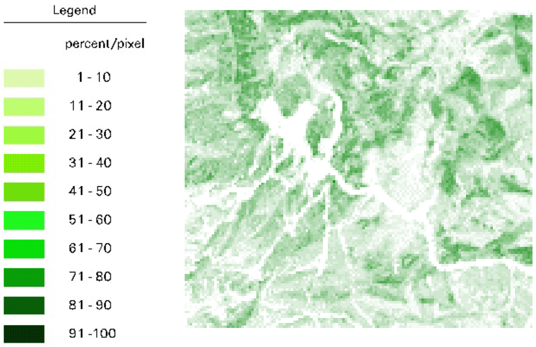


Figure 6. Tree Canopy results, Snow Basin, Utah (2002 Winter Olympic Downhill Ski location).

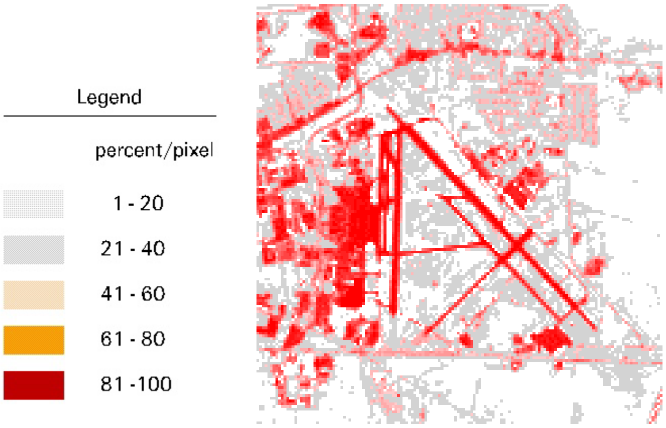


Figure 7. Imperviousness results, Richmond, Va. airport vicinity.

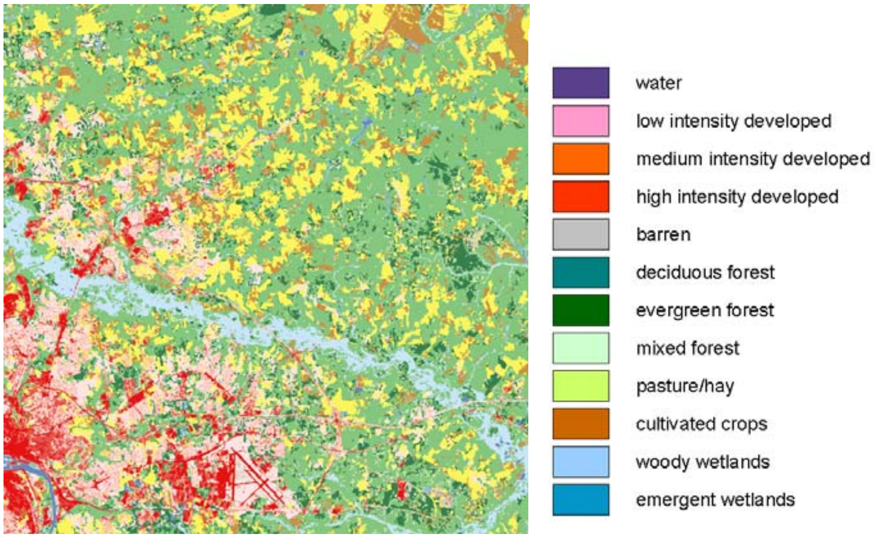


Figure 8. Land cover results, Richmond, Va vicinity.

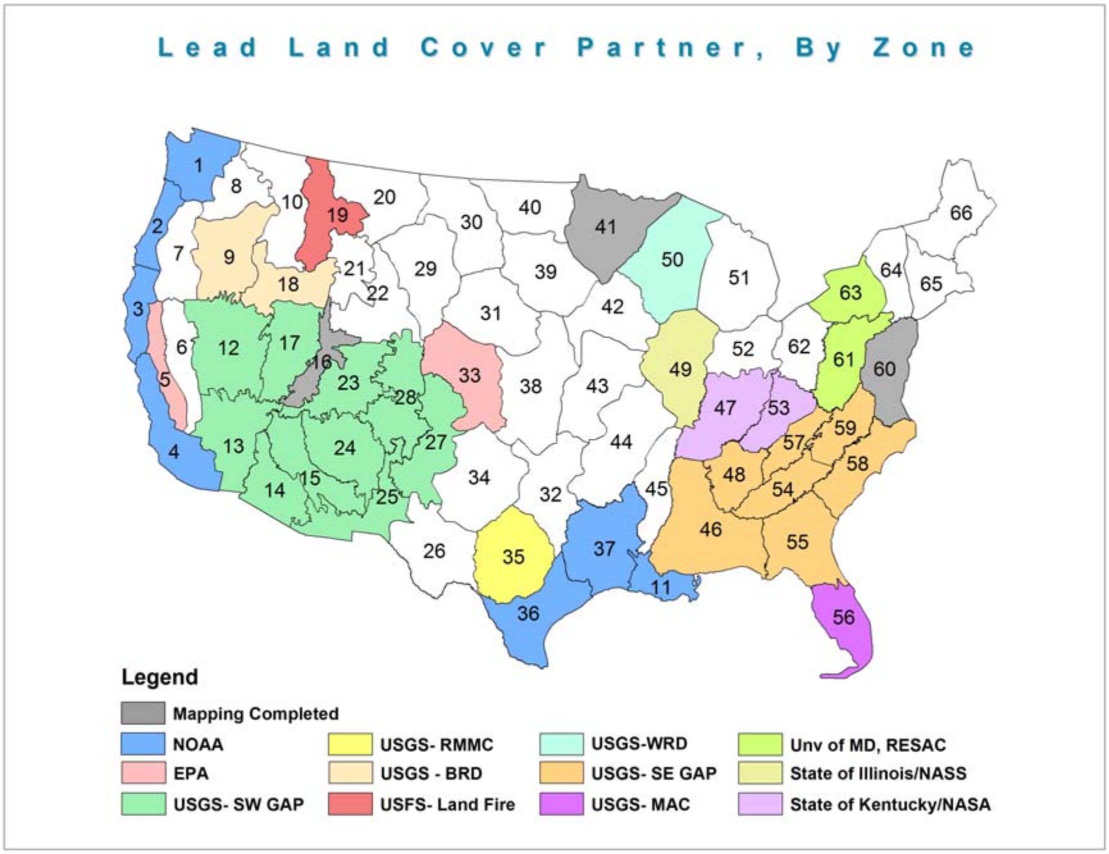


Figure 9. NLCD 2001 current major partners, by mapping zone.

Table 1. NLCD 2001 Land Cover Class Descriptions

11. **Open Water** – All areas of open water, generally with less than 25% cover of vegetation or soil.

12. **Perennial Ice/Snow** – All areas characterized by a perennial cover of ice and/or snow, generally greater than 25% of total cover.

21. **Developed, Open Space -** Includes areas with a mixture of some constructed materials, but mostly vegetation in the form of lawn grasses. Impervious surfaces account for less than 20 percent of total cover. These areas most commonly include large-lot single-family housing units, parks, golf courses, and vegetation planted in developed settings for recreation, erosion control, or aesthetic purposes

22. **Developed, Low Intensity** - Includes areas with a mixture of constructed materials and vegetation. Impervious surfaces account for 20-49 percent of total cover. These areas most commonly include single-family housing units.

23. **Developed, Medium Intensity -** Includes areas with a mixture of constructed materials and vegetation. Impervious surfaces account for 50-79 percent of the total cover. These areas most commonly include single-family housing units.

24. **Developed, High Intensity** - Includes highly developed areas where people reside or work in high numbers. Examples include apartment complexes, row houses and commercial/industrial. Impervious surfaces account for 80 to100 percent of the total cover.

31. **Barren Land** *(*Rock/Sand/Clay*)* - Barren areas of bedrock, desert pavement, scarps, talus, slides, volcanic material, glacial debris, sand dunes, strip mines, gravel pits and other accumulations of earthen material. Generally, vegetation accounts for less than 15% of total cover.

32. **Unconsolidated Shore* -** Unconsolidated material such as silt, sand, or gravel that is subject to inundation and redistribution due to the action of water. Characterized by substrates lacking vegetation except for pioneering plants that become established during brief periods when growing conditions are favorable. Erosion and deposition by waves and currents produce a number of landforms representing this class.

41. **Deciduous Forest** - Areas dominated by trees generally greater than 5 meters tall, and greater than 20% of total vegetation cover. More than 75 percent of the tree species shed foliage simultaneously in response to seasonal change.

42. **Evergreen Forest** - Areas dominated by trees generally greater than 5 meters tall, and greater than 20% of total vegetation cover. More than 75 percent of the tree species maintain their leaves all year. Canopy is never without green foliage.

43. **Mixed Forest** - Areas dominated by trees generally greater than 5 meters tall, and greater than 20% of total vegetation cover. Neither deciduous nor evergreen species are greater than 75 percent of total tree cover.

51. **Dwarf Scrub** – Alaska only areas dominated by shrubs less than 20 centimeters tall with shrub canopy typically greater than 20% of total vegetation. This type is often co-associated with grasses, sedges, herbs, and non-vascular vegetation.

52. **Shrub/Scrub** - Areas dominated by shrubs; less than 5 meters tall with shrub canopy typically greater than 20% of total vegetation. This class includes true shrubs, young trees in an early successional stage or trees stunted from environmental conditions.

71. **Grassland/Herbaceous** - Areas dominated by grammanoid or herbaceous vegetation, generally greater than 80% of total vegetation. These areas are not subject to intensive management such as tilling, but can be utilized for grazing.

72. **Sedge/Herbaceous** – Alaska only areas dominated by sedges and forbs, generally greater than 80% of total vegetation. This type can occur with significant other grasses or other grass like plants, and includes sedge tundra, and sedge tussock tundra.

73. **Lichens** – Alaska only areas dominated by fruticose or foliose lichens generally greater than 80% of total vegetation.

74. **Moss**- Alaska only areas dominated by mosses, generally greater than 80% of total vegetation.

81. **Pasture/Hay** - Areas of grasses, legumes, or grass-legume mixtures planted for livestock grazing or the production of seed or hay crops, typically on a perennial cycle. Pasture/hay vegetation accounts for greater than 20 percent 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 percent of total vegetation. This class also includes all land being actively tilled.

90. **Woody Wetlands** - Areas where forest or shrubland vegetation accounts for greater than 20 percent of vegetative cover and the soil or substrate is periodically saturated with or covered with water.

    91. **Palustrine Forested Wetland* -**Includes all tidal and non-tidal wetlands dominated by woody vegetation greater than or equal to 5 meters in height and all such wetlands that occur in tidal areas in which salinity due to ocean-derived salts is below 0.5 percent. Total vegetation coverage is greater than 20 percent.

    92. **Palustrine Scrub/Shrub Wetland*** - Includes all tidal and non-tidal wetlands dominated by woody vegetation less than 5 meters in height, and all such wetlands that occur in tidal areas in which salinity due to ocean-derived salts is below 0.5 percent. Total vegetation coverage is greater than 20 percent. The

species present could be true shrubs, young trees and shrubs or trees that are small or stunted due to environmental conditions.

93. **Estuarine Forested Wetland*** - Includes all tidal wetlands dominated by woody vegetation greater than or equal to 5 meters in height, and all such wetlands that occur in tidal areas in which salinity due to ocean-derived salts is equal to or greater than 0.5 percent. Total vegetation coverage is greater than 20 percent.

94. **Estuarine Scrub/Shrub Wetland*** - Includes all tidal wetlands dominated by woody vegetation less than 5 meters in height, and all such wetlands that occur in tidal areas in which salinity due to ocean-derived salts is equal to or greater than 0.5 percent. Total vegetation coverage is greater than 20 percent.

95. **Emergent Herbaceous Wetlands** - Areas where perennial herbaceous vegetation accounts for greater than 80 percent of vegetative cover and the soil or substrate is periodically saturated with or covered with water.

96. **Palustrine Emergent Wetland (Persistent)* -** Includes all tidal and non-tidal wetlands dominated by persistent emergent vascular plants, emergent mosses or lichens, and all such wetlands that occur in tidal areas in which salinity due to ocean-derived salts is below 0.5 percent. Plants generally remain standing until the next growing season.

97. **Estuarine Emergent Wetland*** - Includes all tidal wetlands dominated by erect, rooted, herbaceous hydrophytes (excluding mosses and lichens) and all such wetlands that occur in tidal areas in which salinity due to ocean-derived salts is equal to or greater than 0.5 percent and that are present for most of the growing season in most years. Perennial plants usually dominate these wetlands.

98. **Palustrine Aquatic Bed* -** The Palustrine Aquatic Bed class includes tidal and nontidal wetlands and deepwater habitats in which salinity due to ocean-derived salts is below 0.5 percent and which are dominated by plants that grow and form a continuous cover principally on or at the surface of the water. These include algal mats, detached floating mats, and rooted vascular plant assemblages.

99. **Estuarine Aquatic Bed* -** Includes tidal wetlands and deepwater habitats in which salinity due to ocean-derived salts is equal to or greater than 0.5 percent and which are dominated by plants that grow and form a continuous cover principally on or at the surface of the water. These include algal mats, kelp beds, and rooted vascular plant assemblages.

*** Coastal NLCD class only**

Table 2. Cross validation results for imperviousness and canopy mapping, by zone.

Mapping Zone	Tree Canopy Mean Absolute Error	Tree Canopy Correlation Coefficient	Imperviousness Mean Absolute Error	Imperviousness Correlation Coefficient
16	9.9	.88	7	.89
41	14.1	.78	4.6	.83
60	8.4	.93	6	.91

Table 3. Cross validation results for land cover, by zone.

Mapping Zone	Standard Error	Overall Accuracy
16	1.2%	70.5%
41	2.1%	72.6%
60	1.2%	77.2%