Landscape Scale Mapping of Forest Inventory Data by Nearest Neighbor Classification

Andrew Lister[1]

Abstract.—One of the goals of the Forest Service, U.S. Department of Agriculture's Forest Inventory and Analysis (FIA) program is large-area mapping. FIA scientists have tried many methods in the past, including geostatistical methods, linear modeling, nonlinear modeling, and simple choropleth and dot maps. Mapping methods that require individual model-based maps to be produced are time and labor intensive. FIA needs a method that will enable efficient production of large numbers of landscape-scale maps for State reports. The current study presents a case study for the State of Ohio of a nearest neighbor classification method that uses multivariate similarity as a criterion for attaching FIA plot data to pixels with unknown forest attributes to make continuous maps of any FIA attribute. The goal of the study was to devise a landscape scale mapping method that could be easily implemented at a national scale.

Introduction

Landscape scale maps of forest attributes have been of interest in the United States for decades. Sargent (1884) produced what appear to be the first relatively detailed, national maps of forestry data, which included timber volume and species group distribution. During the early to mid-20th century, some national-scale and many finer scale forest vegetation mapping efforts were undertaken by Federal, State, and academic researchers (e.g., Braun 1950; EPA 1994; Little 1971, 1981; Shantz and Zon 1924). Most of these data sets were created by either manually or semimanually digitizing vegetation polygons from photos or from field reconnaissance. During the 1980s and early 1990s, however, advanced computers, such as Geographic Information Systems (GIS) and satellite imagery, became more widely available, leading to more sophisticated vegetation mapping efforts (e.g., Zhu 1994, Zhu and Evans 1994). During the 1990s and early 2000s, remote sensing technology, spatial modeling procedures, and statistical software led to further advances in mapping.

Satellite imagery distribution systems, along with the integration of statistical methods for image classification and spatial modeling with GIS, have led to numerous applications of the use of ground inventory data for mapping forest vegetation, as described by Fassnacht *et al.* (2006) and Andersen (1998). The Forest Service's Forest Inventory and Analysis (FIA) program has used its inventory plot data in conjunction with remotely sensed data for mapping for many years (e.g., Frescino *et al*. 2001, Lister *et al*. 2000, McRoberts *et al.* 2002, Moisen and Edwards 1999). Many of these and other techniques, such as linear modeling methods, are not suitable for production-level mapping because a separate model, and its associated overhead (disk storage, processing time, etc.), is generated for each map. A goal of the FIA program is to develop a production-level mapping procedure that efficiently uses staff, computing resources, and time in order to meet its mapping goals (USDA 1998).

Generally, FIA's mapping goals involve creating accurate maps depicting the spatial distribution and levels of forest resources across the landscape. These maps are often included in publications, on Web sites, and in presentations and are meant to support other data that show the quantity, distribution, and health of the Nation's forests. FIA currently uses maps produced for State reports more as graphics and less as GIS data sets. Interest is growing, however, in using FIA maps as geospatial data sets. For example, FIA-based maps were used by the Forest Service's Forest Health Protection program as ancillary inputs to create forest pest risk maps (Downing n.d.).

[1] Research Forester, U.S. Department of Agriculture, Forest Service, Northern Research Station, Northern Monitoring Program/Forest Inventory and Analysis Unit, 11 Campus Boulevard, Suite 200, Newtown Square, PA 19073. E-mail: alister@fs.fed.us.

The advent of a steady stream of imagery data from the Moderate Resolution Imaging Spectroradiometer (MODIS) sensor (Justice *et al.* 1998) has led to increased use of these data for land cover classification. For example, FIA data have recently been used in conjunction with MODIS data and other GIS layers in the creation of two national-scale maps: dry aboveground biomass (Moisen n.d.) and forest type group (Ruefenacht n.d.). These maps were made using classification and regression trees (Breiman *et al.* 1984) and required large amounts of time and effort to produce.

A more efficient alternative for FIA's future national mapping needs is an approach based on supervised classification. In supervised classification, representative data (here, MODIS imagery and other GIS predictor data linked spatially with the FIA plot information) are used as a reference set. Pixels with unknown values for FIA attributes are populated with "nearest neighbor" reference data based on multivariate similarity between vectors of predictor data at these known locations and those at unknown locations. FIA data have been used in this way before, but generally over smaller areas, with Landsat data or for fewer attributes (e.g., McRoberts *et al.* 2002). In the current study, I present a case study of a nearest-neighbor, supervised classification method that uses FIA data, MODIS imagery, and GIS layers to create landscape maps of the State of Ohio. My goals are to produce maps that will be used for the FIA report that describes Ohio's forests and to present a framework for a methodology that can be applied to other large-area mapping problems.

Methods

Data from 691 homogeneous (single-condition) forested FIA plots collected in Ohio (fig. 1a) between 2001 and 2006 were used in the study. I assume here that the nominal date of the plot information is 2001 and that only marginal changes occurred between the date of the imagery used in the study (2001) and the date the last plot was measured (2006). The distribution of plots in the study area is based on a hexagonal tessellation with one FIA plot randomly located within each 6,000 acre (2,428 ha) hexagon. Each FIA plot consists of four circular 48-ft (14.6-m) diameter subplots, with one subplot located in the center and three equidistant subplots distributed symmetrically around and located 120 ft (31.6 m) from the center subplot. The subplots occupy 0.17 acre (0.07 ha), and the subplot array can be subtended by a circle of 1.5 acre (0.6 ha) in area. FIA attributes summarized to the plot level include basal area per acre of red maple (BARM) found on the plots, cubic foot gross volume (CFGV) of trees greater than 5 in (12.7 cm) diameter at breast height, occurrence of mixed upland hardwoods forest type (MUH), and total dry aboveground biomass (TDRYBIO). For details on the FIA plot design, sample layout, and statistical analytical methods, see Bechtold and Patterson (2006).

The predictor data used were contained in a multilayered Erdas IMAGINE image and consisted of 271 250-m resolution layers, including multidate and monthly composites and derived indices of imagery from the MODIS satellite borne sensor (Justice *et al.* 1998), several rasterized summaries of the STATSGO soils database compiled by the Natural Resources Conservation Service (1994), summaries of the landcover classes found in the U.S. Geological Survey National Land Cover Dataset (NLCD) database (Vogelmann *et al.* 2001), mean monthly and annual temperature and precipitation from the PRISM climate database (Daly *et al.* 2004), a rasterized grid representing distance to streams (USGS 1999), and various derivatives of the National Elevation Dataset (Gesch *et al.* 2002). Complete details of the steps used to prepare the data and data derivatives are on file at the Forest Service's Northern Research Station (11 Campus Blvd, Ste. 200, Newtown Square, PA 19073). These data sets were precompiled, mosaicked, and clipped to U.S. Geological Survey NLCD 2001 mapping zones (which are similar to ecoregions) (Homer and Gallant 2001) for the United States, and portions of the zones that intersect Ohio (zones 62, 53, 52, 51, and 47) were mosaicked to create a predictor data set.

The feature selection method I used was meant to find a subset of the predictor data set that would be effective at discriminating plots that are ecologically different from one another. The assumption of doing this is that plots have a unique ecological signature in feature space that can be used as a reference data set for labeling unknown locations in feature space. In order to select an effective subset of the predictor data set, a three-step process was used. The steps of this process were to (1) classify the plots based on the species composition data (independent of

the GIS predictor data), (2) rank the predictor attributes based on their ability to discriminate this species composition class, and then (3) choose a subset of predictors based on this ranking. First, WEKA[2] data mining software was used to classify each plot into one of 10 species composition classes (hereafter referred to as "forest types") using k-means cluster analysis (Witten and Frank 2005), with the total basal area of each species forming the axes used for clustering. In other words, forest types were created by forming 10 classes, each of which contained plots with similar species composition. I arbitrarily chose 10 classes because exploratory analyses showed that choosing around 10 yielded the most even distribution of points across the clusters.

Next, WEKA's implementation of the RELIEF attribute selector (Kira and Rendell 1992) was used to rank each GIS predictor variable. RELIEF worked by creating a usefulness index for each predictor attribute by randomly choosing a large number of instances from the data set and calculating for each selected instance the difference between the GIS predictor variable's value for the closest instance of the same "forest type" and that of the closest member of a different forest type. In the current study, if instances that were located close together along the axis defined by a predictor attribute are of the same forest type, RELIEF considered the attribute useful for discriminating between plots with different ecological characteristics and ranked it higher.

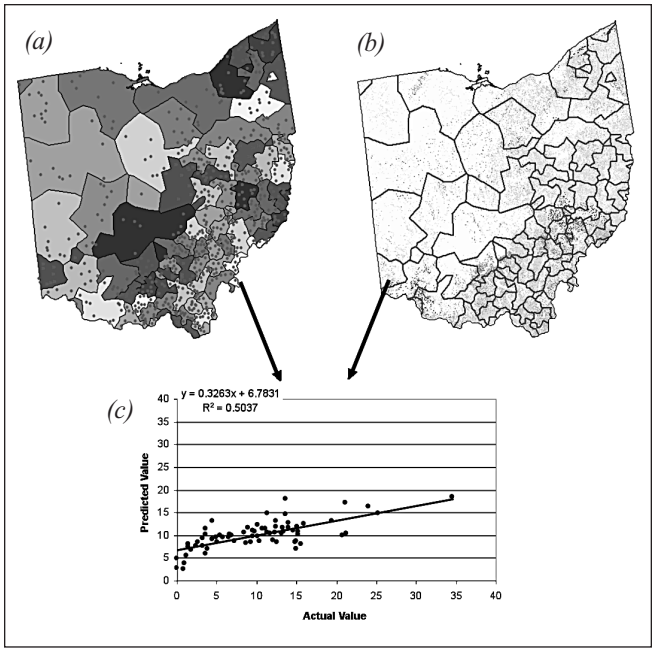
Finally, using the results of the attribute ranking, an arbitrary assessment of Pearson correlation matrices (to eliminate collinear variables with similar ranks), and my judgment from past use of similar data, I subjectively chose a set of 13 predictors that was useful for discriminating species composition class and was minimally correlated (correlations are generally less than 0.25). The 13 variables chosen in this manner were, in order of usefulness, MODIS band 5 (May 9, 2001), MODIS Enhanced Vegetation Index (EVI) (August 14, 2001), NLCD percent woody wetland, MODIS band 3 (November 17, 2001), distance to streams, the count of the variety of aspect values derived from a digital elevation model (an index of topographic roughness), MODIS band 7 (April 7, 2001), soil pH, soil texture, X coordinate, rock volume, Y coordinate, and minimum temperature in November.

Leica Geosystems' IMAGINE[2] image processing software was used to standardize the data set to the same measurement scale (0-1) and to extract values for each of the 13 standardized predictor layers where the FIA plots used in the analysis were located. IMAGINE's minimum (Euclidean) distance classifier was used to impute FIA plot information from the set of known pixels to unknown pixels based on multivariate similarity. The classification procedure gave every pixel in the study area the plot identifier (id) value of the FIA plot that is most similar to it with respect to values of the 13 predictors. A simple lookup table was then used to link pixel values in the image to tabular plot level summaries of FIA attributes based on this plot id value (as described in Lister (2005)). For this study, the plot id map was recoded to create maps of several attributes: BARM, CFGV, MUH and TDRYBIO.

Quality assurance (QA) was performed by a novel method of grouping plots into contiguous clusters, with each cluster containing exactly 10 plots (fig. 1a). This grouping was done by bit interleaving of the x and y coordinates of the plots in a manner similar to that described by Faloutsos and Rong (1991). Bit interleaving can be used to order plots based on proximity in two dimensions by drawing a line with fractal properties through the study area so that each plot is visited exactly once. The fractal line has the property of folding in upon itself in an orderly manner so that, in general, groups of points next to each other on the line are close in space. By partitioning this line into clusters that contain exactly 10 plots each, the procedure tessellates the study area based on density of forested plots. I chose 10 plots for each cluster arbitrarily, based on previous work that found this number to be a reasonable trade off between having a sufficiently small spatial cluster as the analysis unit and a sufficiently large group of plots to characterize that spatial cluster. If the spatial clusters were too big, the analysis became less meaningful, but if not enough plots were in each cluster, the variance of the cluster estimates made the analysis suspect.

By creating a raster representation of the study area and assigning each pixel the cluster id of the closest plot to that pixel, it was possible to summarize both the FIA plot data (fig. 1a) and the mapped estimates (fig. 1b) by cluster and construct a set of simple scatterplots (fig. 1c) that depict the relationship between the actual values (the average plot value) and the

Figure 1.—*(a) Contiguous clusters containing 10 plots each are created. Regions are built around each cluster using a GIS, and cluster-level summaries of the FIA plot data are calculated. (b) Map-based estimates are summarized for each region. (c) Scatterplots are created along with simple linear regression line diagnostics to characterize the actual vs. predicted relationship.*



estimates (the average pixel value). These relationships were then characterized by the parameters of the simple linear regression line (slope, intercept, and R^2) that describes the relationship between the set of actual and predicted values. The regression line method I used is a descriptive technique meant simply as a QA tool—I did not attempt to make inferences about the significance of the parameter estimates. The goal of the QA procedure was to provide a tool for data consumers to assess the relationship between actual and predicted values in a spatially explicit manner.

Results and Discussion

Many of the attributes that were selected by the RELIEF method are factors that would tend to influence vegetation composition in a landscape dominated by the effects of past glaciations, like Ohio. For example, soil pH, texture, and rock volume would be affected by past glacial activity, as would stream density and wetland occurrence. The MODIS-related imagery variables probably appeared higher in the usefulness ranking because of differences in phenology of the different species composition groups—certain species assemblages reflect light differently at different times of the year. The goal of the attribute selection approach was not strictly to extract biologically meaningful predictor variables in a quantitative way. Rather, it was to use a combination of RELIEF, correlation tests, and user opinion as a guide in attribute selection. It is noteworthy that the chosen variables (which were ranked higher by RELIEF) probably have at least some functional relationships with factors that affect plant growth in Ohio.

The maps of the FIA attributes selected are shown in figure 2 along with magnified areas to show examples of the finer scale variability of the estimates. Ohio is nearly 70 percent nonforest, so I used a nonforest mask (the NLCD 1992 data) to mask out nonforest areas and water, which accounts for some of the observed patterns in the maps. Nonforest masking allows the map to retain certain landscape features that FIA doesn't measure (e.g., the occurrence of rivers) while imputing FIA attributes to forested areas of the State.

In general, the southwest part of the State has more biomass and volume than other parts of the State. The landscape maps produced clearly show these patterns of the FIA attributes across the landscape, and the example areas that are shown at higher resolution show some of the finer scale pattern that the modeling procedure produces. In general, however, the interpretation of these maps is best made at the landscape level. Pixel-scale interpretation is possible, although inadvisable because it is nearly impossible to find QA reference data that correspond well with MODIS pixels. The FIA plots cover 0.067 ha, or approximately 1/100 of a 250-m MODIS pixel—thus making plot-pixel accuracy statements nearly meaningless.

On the other hand, the QA results within the zones depicted in figure 1 can serve to inform the user about the relative utility of the maps at a given geographic scale. Figures 3a, 3b, 3c, and 3d show the relationships (and corresponding diagnostic statistics) between sets of actual and predicted levels of the FIA attributes. The r^2 values ranged from 0.4 (for the biomass-related

Figure 2.—*Maps of total dry aboveground biomass (TDRYBIO), occurrence of mixed hardwoods forest type (MUH), cubic foot gross volume (CFGV) of trees, and basal area per acre of red maple (BARM), with small areas shown at higher resolution to show local detail.*

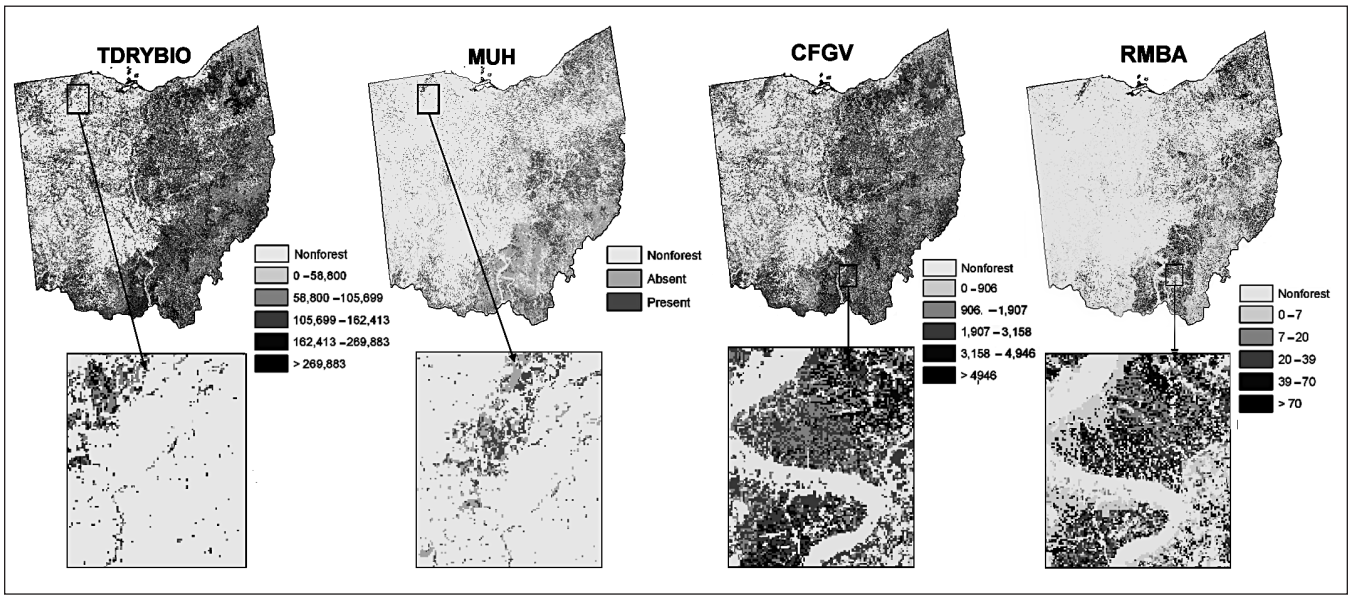
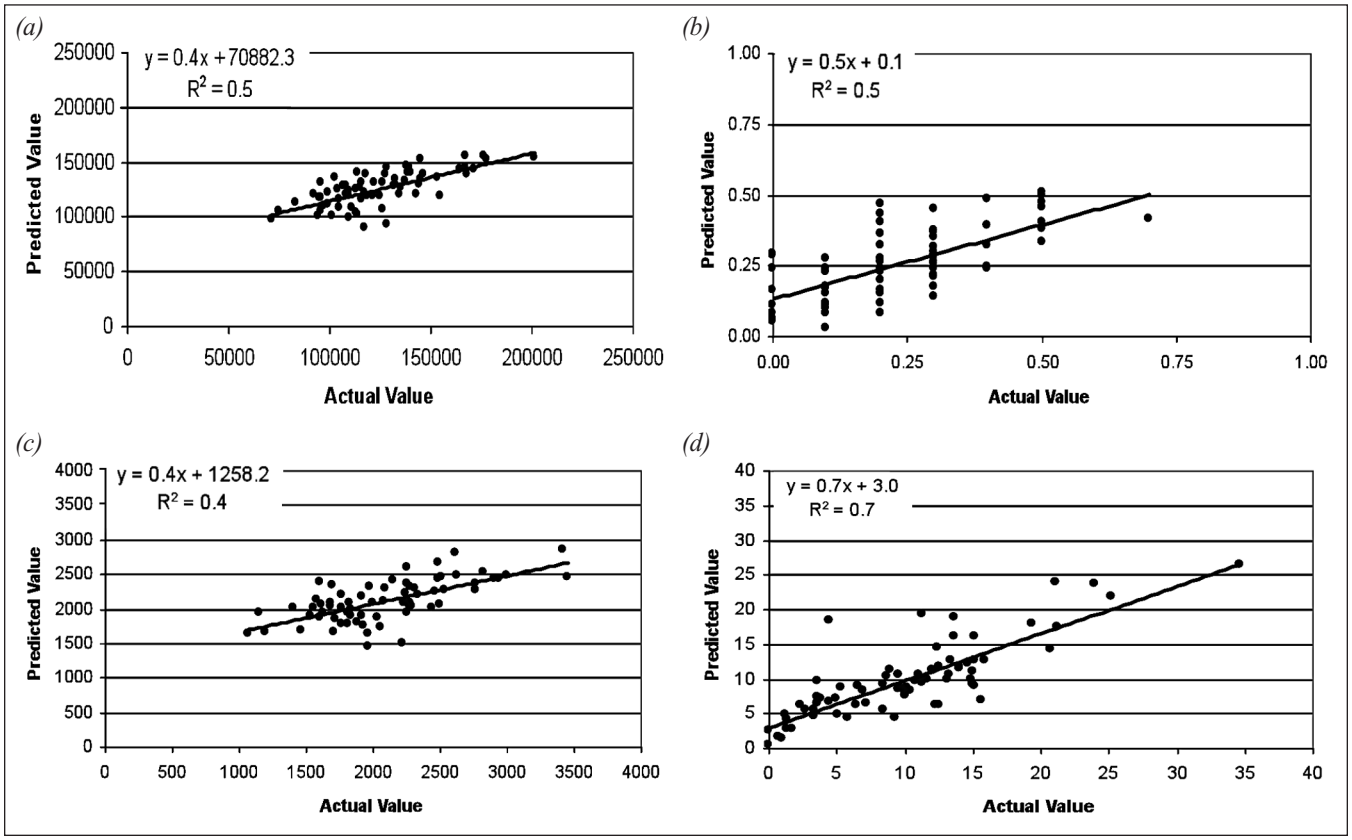


Figure 3.—*Actual vs. predicted scatterplots and associated simple linear regression diagnostics of (a) total dry aboveground biomass, (b) mixed upland hardwoods forest type, (c) cubic foot gross volume, and (d) basal area per acre of red maple.*



variables) to 0.7 (for the basal of red maple). The slope and intercepts of the simple linear regression lines that describe the relationships between actual and predicted values tend to indicate an overprediction of low actual values and an underprediction of high values. I have encountered this phenomenon in many other multivariate modeling methods and believe that it occurs because weak univariate and multivariate relationships between an FIA plot and the MODIS pixel on which it sits tend to increase the occurrence of misclassification. In most FIA data sets, a random assignment of an FIA plot to an unknown pixel (which is the extreme of what occurs in misclassification) will always yield an estimate closer to the mean of the value for all FIA plots involved than it will a value near the extremes. This principle leads to the observed pattern of truncation of the variance of the set of estimates.

The main reason I chose the novel QA method I used was to guarantee that an equal number of FIA plots would be in each QA zone (fig. 1a). In that manner, the confidence I can put in each QA point is equal with respect to the FIA plots, which generally show the largest amount of variability. Had I chosen another approach using a regular tessellation of the study area to produce QA zones, large areas of the State would not have been assessed because cells in mostly nonforest areas would not have at least 10 FIA plots and would not be used as valid QA polygons. By grouping plots using the bit interleaving method, not only am I able to perform the QA method within contiguous geographic regions but I also have increased the interpretability of the results of the analysis. Interpreting these QA results is predicated on recognizing that each QA zone represents a roughly equal amount of *forest* land, not total land.

The utility of landscape maps such as these is tempered by the truncation of the variance I observe in the QA results. Were the goal of my study to create an accurate map of a single FIA attribute, I would have optimized my choice of predictors and modeling technique. For the purposes of FIA's State reporting, however, a method that produces a single map (the plot id map) and uses a lookup table to create a map of any FIA attribute that can be associated with a plot is clearly desirable. The landscape maps that I have produced not only show the distribution of the attributes of interest across the landscape but also retain logical consistency. Each map produced retains the entire plot record for each pixel; thus, e.g., a situation where the basal area of red maple at a given location is predicted to be higher than that of the total basal area of all species cannot occur. The disadvantage of this technique, however, is that any individual map that was not produced using optimal methods or predictor data is not optimized for accuracy. New imputation techniques, like those being implemented by Wilson (2006), use advanced data reduction methods and attribute weighting, which could mitigate this problem. These advanced methods show great promise for national implementation, and future work will be in support of this goal.

Literature Cited

Andersen, G.L. 1998. Classification and estimation of forest and vegetation variables in optical high resolution satellites: a review of methodologies. Interim report IR-98-085. Laxenburg, Austria: International Institute for Applied Systems Analysis. 25 p.

Bechtold, W.A.; Patterson, P.L. 2005. The enhanced Forest Inventory and Analysis program—national sampling design and estimation procedures. Gen. Tech. Rep. SRS-80. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 85 p.

Braun, E.L. 1950. Deciduous forests of eastern North America. Philadelphia, PA: Blaskiston Company. 596 p.

Breiman, L.; Friedman, J.; Olshen, R.A.; Stone, C.J. 1984. Classification and regression trees. Pacific Grove, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Daly, C.; Gibson, W.P.; Doggett, M.; Smith, J.; Taylor, G. 2004. Up-to-date monthly climate maps for the conterminous United States. In: Proceedings, 14th American Meteorological Society conference on applied climatology, 84th American Meteorological Society annual meeting combined preprints. Pap. P5.13. Seattle, WA: American Meteorological Society. http://ams.confex.com/ams/pdfpapers/71444.pdf.

Downing, M. [N.d.]. Invasive pest risk maps. On file with: M. Downing, Forest Health Protection/Forest Health Technology Enterprise Team, 2150 Centre Avenue, Building A, Suite 331, Information Technology, Fort Collins, CO 80526–1891. http://www.fs.fed.us/foresthealth/technology/invasives_sirexnoctilio_riskmaps.shtml.

Faloutsos, C.; Rong, Y. 1991. DOT: a spatial access method using fractals. In: Proceedings of the International Conference on Data Engineering (ICDE). Kobe, Japan: Institute of Electrical and Electronics Engineers: 152-159.

Fassnacht, K.S.; Cohen, W.B.; Spies, T.A. 2006. Key issues in making and using satellite-based maps in ecology: a primer. Forest Ecology and Management. 222: 167-181.

Frescino, T.S.; Edwards, T.C., Jr.; Moisen, G.G. 2001. Modeling spatially explicit forest structural attributes using generalized additive models. Journal of Vegetation Science. 12: 15-26.

Gesch, D.; Oimoen, M.; Greenlee, S.; Nelson, C.; Steuck, M.; Tyler, D. 2002. The national elevation dataset. Photogrammetric Engineering & Remote Sensing. 68(1): 5-32.

Homer, C.; Gallant, A. 2001. Partitioning the conterminous United States in mapping zones for Landsat TM land cover mapping. Sioux Falls, SD: U.S. Department of the Interior, U.S. Geological Survey, Eros Data Center. 7 p.

Justice, C.; Vermote, E.; Townshend, J.R.G. *et al.* 1998. The Moderate Resolution Imaging Spectroradiometer (MODIS): Land remote sensing for global change research. IEEE Transactions on Geoscience and Remote Sensing. 36(4): 1228-1249.

Kira, K.; Rendell, L.A. 1992. A practical approach to feature selection. Proceedings, ninth international conference on machine learning. Aberdeen, Scotland: Morgan Kaufmann: 249-256.

Lister, A.J. 2005. Creation of an n-dimensional spatial database instead of a map. In: Marsden, M.; Downing, M.; Riffe, M., comps. Workshop proceedings: quantitative techniques for deriving national-scale data. FHTET-05-12. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Forest Health Technology Enterprise Team: 14-22.

Lister, A.J.; Riemann, R.; Hoppus, M. 2000. A nonparametric geostatistical approach for estimating species importance. In: Reams, G.A.; McRoberts, R.E.; VanDeusen, P.C., eds. Proceedings, second annual forest inventory and analysis symposium. USDA GTR SRS-47. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station: 52-62.

Little, E.L., Jr. 1971. Atlas of United States trees. Misc. Publ. 1146. Washington, DC: U.S. Department of Agriculture. 9 p. Vol. 1.

Little, E.L., Jr. 1981. Atlas of United States trees. Supplement. Misc. Publ. 1410. Washington, DC: U.S. Department of Agriculture. 31 p. Vol. 5.

McRoberts, R.E.; Nelson, M.D.; Wendt, D.G. 2002. Stratified estimation of forest area using satellite imagery, inventory data, and the *k*-Nearest Neighbors technique. Remote Sensing of Environment. 82(2): 457-468.

Moisen, G. [N.d.]. Forest biomass of the conterminous United States. Manuscript in preparation. On file with: G. Moisen, U.S. Department of Agriculture, Forest Service, Forest Inventory and Analysis Unit, Rocky Mountain Research Station, 507 25th Street, Ogden, UT 84401.

Moisen, G.G.; Edwards, T.C., Jr. 1999. Use of generalized linear models and digital data in a forest inventory of northern Utah. Journal of Agricultural, Biological, and Environmental Statistics. 4: 372-390.

Natural Resources Conservation Service (NRCS). 1994. State Soil Geographic (STATSGO) soil database: data use information. Misc. Publ. 1492. Washington, DC: National Soil Survey Center. 88 p.

Ruefenacht, B. [N.d.]. Forest type groups of the conterminous United States. Manuscript in preparation. On file with: B. Ruefenacht, U.S. Department of Agriculture, Forest Service, Remote Sensing Applications Center, Salt Lake City, UT 84119. http://svinetfc4.fs.fed.us/rastergateway/forest_type/.

Sargent, S. 1884. Report on the forests of North America (exclusive of Mexico). Washington, DC: U.S. Department of the Interior, Government Printing Office.

Shantz, H.L.; Zon, R. 1924. Atlas of American agriculture. Washington, DC: U.S. Department of Agriculture, Government Printing Office.

U.S. Department of Agriculture (USDA). 1998. A Forest Service Forest Inventory and Analysis (FIA) strategic plan for forest inventory and monitoring. Washington, DC: U.S. Department of Agriculture, Forest Service. 48 p.

U.S. Environmental Protection Agency (EPA). 1994. GIRAS Landuse/Landcover data for the conterminous United States by quadrangles at scale 1:250,000. Washington, DC: U.S. Environmental Protection Agency, Office of Information Resources Management.

U.S. Geological Survey (USGS). 1999. The National Hydrography Dataset. Fact Sheet 106-99. Reston, VA: U.S. Geological Survey. 2 p.

Vogelmann, J.E.; Howard, S.M.; Yang, L. *et al.* 2001. Completion of the 1990s National Land Cover Data Set for the conterminous United States from Landsat Thematic Mapper data and ancillary data sources. Photogrammetric Engineering and Remote Sensing. 67(6): 650-662.

Wilson, B.T. 2006. Northern Research Station FIA program's atlas of FIA maps. Unpublished document. On file with: B.T. Wilson, Northern Research Station, Forest Inventory and Analysis program, St. Paul, MN 55108.

Witten, I.H.; Eibe, F. 2005. Data mining: practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann. 525 p.

Zhu, Z. 1994. Forest density mapping in the lower 48 States: a regression procedure. Res. Pap. SO-280. New Orleans: U.S. Department of Agriculture, Forest Service, Southern Forest Experiment Station. 11 p.

Zhu, Z.; Evans, D.L. 1994. U.S. forest types and predicted percent forest cover from AVHRR data. Photogrammetric Engineering and Remote Sensing. 60(5): 525-531.