SYNERGISTIC USE OF FIA PLOT DATA AND LANDSAT 7 ETM+ IMAGES FOR LARGE AREA FOREST MAPPING

Chengquan Huang[1], Limin Yang[1], Collin Homer[1], Michael Coan[1], Russell Rykhus[1], Zheng Zhang[1], Bruce Wylie[1], Kent Hegge[1], Zhiliang Zhu[2], Andrew Lister[3], Michael Hoppus[3], Ronald Tymcio[4], Larry DeBlander[4], William Cooke[5], Ronald McRoberts[6], Daniel Wendt[7], and Dale Weyermann[8]

ABSTRACT.—FIA plot data were used to assist in classifying forest land cover from Landsat imagery and relevant ancillary data in two regions of the U.S.: one around the Chesapeake Bay area and the other around Utah. The overall accuracies for the forest/nonforest classification were over 90 percent and about 80 percent, respectively, in the two regions. The accuracies for deciduous/evergreen/mixed and forest type group classifications were around 80 percent and 65 percent, respectively, and were consistent in the two regions. These results suggest that use of FIA plot data together with satellite imagery and relevant ancillary data may substantially improve the efficiency, accuracy, and consistency of large area forest land cover mapping.

Reliable and updated forest information is required for many scientific and land management applications. Meeting this requirement is of interest to both the Forest Inventory and Analysis (FIA) program of the USDA Forest Service and the Land Cover Characterization (LCC) Program of the U.S. Geological Survey (USGS) Earth Resources Observation Systems (EROS) Data Center (EDC). FIA has a mandate to collect and report information periodically on status and trends in the Nation's forested resources, while the LCC program has a mandate to develop a circa 2000 national land cover database through the Multi-Resolution Land Characterization (MRLC) 2000 project. Therefore, it is in the best interest of the government that these two agencies collaborate in mapping the Nation's land cover. The current study is the result of an initial collaboration between the two agencies.

Forest land cover information is often derived from remotely sensed images using classification algorithms (e.g., Franklin and others 1986, Mickelson and others 1998), many of which require substantial amount of reference data (Hall and others 1995, Townshend 1992). Reliable reference data are also required for assessing classification results. One of the challenges to mapping forest land cover over large areas is the lack of adequate reference data. In areas where some reference data sets exist, they may have been collected in different ways and may have different levels of reliability. Such scarcity of reliable reference data and lack of consistency among the available data sets often limit the efficiency, consistency, and accuracy in deriving forest information from satellite imagery.

The plot data collected through FIA make up a potentially high quality reference data set for the MRLC 2000 project. FIA plots represent a statistically based sampling of the Nation's land. Detailed information on forest status and

[1] USGS EROS Data Center, Raytheon, Sioux Falls, SD 57198.
[2] USGS EROS Data Center, Sioux Falls, SD 57198.
[3] USDA Forest Service, Northeastern Research Station, 11 Campus Blvd., Suite 200, Newtown Square, PA 19073.
[4] USDA Forest Service, Rocky Mountain Research Station, Ogden, UT 84401.
[5] USDA Forest Service, Southern Research Station, P.O. Box 928, Starkville, MS 39760.
[6] USDA Forest Service, North Central Research Station, 1992 Folwell Ave., St. Paul, MN 55108.
[7] USDA Forest Service, R9, Milwaukee, WI.
[8] USDA Forest Service, Pacific Northwest Research Station, 1221 SW Yamhill, Suite 200, Portland, OR 97205.

structure is collected periodically at each plot through intensive field work. With minimal efforts, this data set can be reorganized for use with remotely sensed images. The purpose of the current study is to evaluate the usefulness of the FIA plot data in deriving forest land cover classifications from satellite imagery over large areas and to test whether using this data set can improve the efficiency, accuracy, and consistency in developing the MRLC 2000 national land cover database.

DATA AND PREPROCESSING

Study Area

For mapping efficiency, the conterminous United States was divided into 66 mapping zones for the MRLC project. Two mapping zones—zones 16 and 60—were used in this pilot study (fig. 1). Zone 60 represents the eastern coastal environment, covering the Chesapeake Bay area, while zone 16 represents the western arid and less developed landscape, covering Utah and southern Idaho. Figure 1 shows the Landsat paths/rows covered by the two mapping zones.

Landsat Imagery and Ancillary Data

For each Landsat path/row covered by the two mapping zones, Enhanced Thematic Mapper Plus (ETM+) images were acquired on three different dates to capture vegetation dynamics over a growing season and to maximize land cover type separability (Yang and others 2001a). These images were acquired between 1999 and 2001 and were selected to minimize the impact of cloud cover and atmospheric effects. The images were geometrically and radiometrically corrected using standard methods at the USGS EROS Data Center (Irish 2000). Terrain correction using the USGS 1-arc second National Elevation Dataset was performed to improve geolocation accuracy. Raw digital numbers were converted to at-satellite reflectance for the six reflective bands and to at-satellite temperature for the thermal band according to Markham and Barker (1986) and the Landsat 7 Science Data User's Handbook (Irish 2000). All seven bands were resampled to a 30-m spatial resolution. Tasseled-cap brightness, greenness, and wetness were calculated using at-satellite reflectance based coefficients (Huang and others 2002b).

Ancillary data included the USGS 1-arc second National Elevation Dataset and three derivatives: slope, aspect, and a topographic position index. In addition, three soil attributes—available water capacity, soil carbon content and a soil quality index—were derived from the State Soil Geographic (STATSGO) Data Base. All ancillary data layers were resampled to a spatial resolution of 30 m.

Reference Data Sets

Through intensive field work, the FIA program provides detailed forest attributes at individual tree, subplot, and plot levels. Considering the pixel size of the ETM+ imagery and

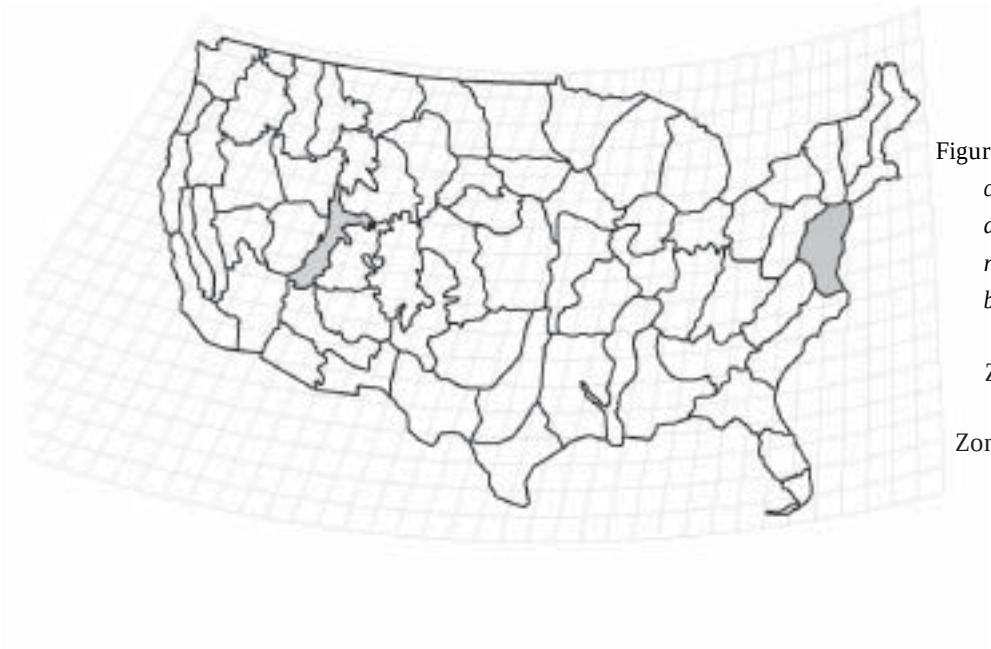

Figure 1.— *Mapping zones of the conterminous U.S., with the two study areas shaded. The background grid represents Landsat 7 path/row boundary.*

possible geolocation errors of the imagery and FIA plots, only plot-level data were deemed appropriate for use with the ETM+ imagery. Therefore, tree-level data were summarized to subplot level and then to plot level. In zone 60, each plot was labeled with single- or multiple-condition primarily depending on if there were only one or multiple land use/land cover types within the plot area. Multiple-condition plots were excluded to minimize the impact of misregistration errors and other possible inconsistencies between FIA plots and the satellite images. This was not deemed necessary in zone 16 because most of the plots were based on the plot design used prior to 1995, which restricts all plots to single-condition. Each eligible plot was then classified at three levels: forest/nonforest, deciduous/evergreen/mixed, and forest type group. Table 1 lists the number of FIA plots used in this study. Global Positioning System units were used to locate all plots in zone 60 and 147 plots in zone 16. The remaining plots in zone 16 were digitized from aerial photos used in the field. Geolocation errors between the digitized plots and ETM+ images were minimal. A visual check of over 100 random plots digitized from the air photos against corresponding satellite images suggested that less than 10 percent of the plots had location errors greater than one ETM+ pixel.

Table 1.—*Number of FIA plots used in this study*

	Zone 60	Zone 16
Forest/nonforest	1,750	3,037
Deciduous/evergreen/mixed	669	1,754
Forest type group	669	1,852

Another reference data set available in zone 16 consisted of field data collected by the Fire Science Lab of the Rocky Mountain Research Station of the USDA Forest Service and the Utah GAP Analysis Program of Utah State University. Each field site was classified at two levels: forest/nonforest and deciduous/evergreen/mixed. This was used as an independent data set to evaluate the classification results developed using FIA plot data in mapping zone 16. Although the point location was not based on any statistically rigorous sampling design and the evaluation should not be considered a statistically rigorous accuracy assessment, this independent reference data set should provide useful information on the consistency of FIA plot data and the reliability of derived classifications.

METHODS

Classification Levels

As with the reference data, classification of the ETM+ images and ancillary data was performed at three levels: forest/nonforest, deciduous/evergreen/mixed, and forest type group. A forest/nonforest map is required by FIA to implement a stratified sampling of forested land in order to produce accurate estimates of forest attributes. Deciduous, evergreen, and mixed are the main forest categories in the MRLC 2000 classification scheme. Type group information is often required for species conservation planning, fire management, and many other applications. Table 2 lists the major forest type groups in the two mapping zones.

Table 2.—*Forest type groups in the two mapping zones*

Zone 60	Zone 16
Spruce/fir	Pinyon/juniper
Loblolly & shortleaf pine	Douglas-fir
Oak/pine	Ponderosa pine
Oak/hickory	Fir/spruce/mountain hemlock
Oak/gum/cypress	Lodgepole pine
Elm/ash/red maple	Other western softwoods
	Aspen/birch
	Western oak
	Other western hardwoods

Decision Tree Classifier

Many algorithms are available for classifying satellite images (Hall and others 1995, Townshend 1992); among the most popular of these include the maximum likelihood classifier, neural network classifiers, and decision tree classifiers. Decision tree was chosen for this study because it 1) is non-parametric and therefore independent of the distribution of class signature, 2) can handle both continuous and nominal variables, 3) generates interpretable classification rules, and 4) is fast to train and is often as accurate as, and sometimes more accurate than, many other classifiers (Hansen and others 1996, Huang and others 2002a). The decision tree program used in this study, C5, employs an information gain ratio criterion in tree development and pruning (Quinlan 1993). This program has many advanced features, including boosting and cross-validation.

Boosting

Boosting is a technique for improving classification accuracy (Bauer and Kohavi 1998). With this function, the program develops a sequence of decision trees, each subsequent one trying to fix the misclassification errors in the previous tree. Each decision tree makes a prediction. The final prediction is a weighted vote of the predictions of all trees. This function often improves classification accuracy by 5 to 10 percent (e.g., Friedl and others 1999).

Cross-Validation

Cross-validation is designed to obtain relatively realistic accuracy estimates using a limited number of reference data samples for both training and accuracy assessment (Michie and others 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 classification developed using the remaining training samples, and their average value represents the accuracy of the classification developed using all reference samples.

RESULTS AND DISCUSSION

Classification accuracies at all three levels in the two mapping zones were estimated through cross-validation (table 3). These accuracy estimates can be considered reasonably realistic, because the FIA plots are not spatially auto-correlated, they cover the entire of each study area, and their locations were determined through statistically based sampling designs (Michie and others 1994). This point is demonstrated by the fact that for the forest/nonforest and deciduous/evergreen/mixed classifications in zone 16, the accuracies estimated using the independent reference data set collected by the Fire Science Lab of the Rocky Mountain Research Station and the Utah GAP Analysis Program were similar to those estimated through cross-validation (table 3).

With the boosting function of the C5 program, overall accuracies of around 80 to 90 percent, 80 percent, and 65 percent were achieved in both mapping zones for the forest/nonforest, deciduous/evergreen/mixed, and forest type group classifications, respectively. At the three classification levels, the boosting function improved classification accuracy by about 2 to 10 percent in absolute values. Similar improvements using the boosting function have been reported in other studies (e.g., Chan and others 2001). The final classifications in the two study areas were developed using the boosting function. The classifications in zone 16 were evaluated by field crews of the Rocky Mountain Research Station and the Utah GAP Analysis Program. Both parties agreed that these classifications were reasonably accurate.

Despite the very different landscapes, classification accuracies for the two mapping zones are comparable at the deciduous/evergreen/mixed level and at the forest type group

Table 3.—*Classification accuracy estimates for the two mapping zones*

Classification level	Forest/nonforest		Deciduous/evergreen/mixed		Forest type group	
	Accuracy	Std. error	Accuracy	Std. error	Accuracy	Std. error
	- - - - - - - - - - -	-Percent-	- - - - - - - - - - -			
Zone 60, cross-validation						
Without boosting	90.7	0.4	74.0	1.4	59.6	1.1
With boosting	93.7	0.7	78.9	0.8	66.1	2.2
Zone 16, cross-validation						
Without boosting	80.4	0.4	78.0	0.4	56.6	0.9
With boosting	82.7	0.4	81.2	0.6	65.8	1.2
Zone 16, use of independent test data set						
Without boosting	75.7	-	75.3	-	-	-
With boosting	79.0	-	83.4	-	-	-

level, suggesting that similar accuracies are likely achievable in other areas using FIA plot data, Landsat 7 imagery, and relevant ancillary data.

However, the forest/nonforest classification in zone 16 is about 10 percent less accurate than in zone 60. This is probably because some forest and natural nonforest vegetation are more difficult to separate both spectrally and physiologically in the arid environment of zone 16. Even from the ground, some field crews recognized that it is sometimes very difficult to separate tall shrubs from sparse short trees without ambiguity. Considering the complex topography and the difficulty in defining forests in zone 16, the accuracies achieved in this zone probably represent the lower end of the accuracies expected in forest/nonforest classifications throughout the Nation.

The development of the classifications in each mapping zone took an experienced person about 3 to 4 months, including pre-processing of the ETM+ images and ancillary data discussed earlier. Our experience from developing the 1992 National Land Cover Dataset (Vogelmann and others 2001) suggests that if the FIA plot data had not been readily available for this study, at least one-third extra time and effort would have had to be devoted to reference data collection. For the MRLC 2000 project, even if some resources are available for reference data collection, the spatial coverage and location of collected reference data points very likely will not be as preferable as the FIA plot data.

The ability of the cross-validation to produce accuracy estimates at the classification stage can be highly valuable to many users of regional classifications, because statistically rigorous accuracy assessment of such classifications can be very expensive and often takes a long time before any accuracy estimate can be derived (Yang and others 2001b, Zhu and others 2000). In order for the cross-validation estimates to be as little biased as possible, however, the reference data should not be spatially auto-correlated and should be collected through a statistically based sampling design (Friedl and others 1999, Michie and others 1994). The FIA plot data make up perhaps one of the few readily available reference data sets for regional applications that meet these criteria.

CONCLUSIONS

1. FIA plot data are useful reference data for mapping forest land cover at regional and national scales. Forest maps developed using this data set, Landsat 7 ETM+ image, and ancillary data in the two mapping zones had overall accuracies of about 80 to 90 percent, 80 percent, and 65 percent at the forest/nonforest, deciduous/evergreen/mixed, and forest type group levels, respectively.
2. Use of FIA plot data as part of the reference data set in the MRLC 2000 project can substantially improve mapping efficiency, accuracy, and consistency. The spatial coverage of the plots and the statistically based sampling design of plot location make it possible to produce reasonably realistic accuracy estimates at the classification stage.
3. The decision tree classifier proves a viable and efficient method for deriving forest classifications over large areas. The boosting function can improve classification accuracy by 2 to 10 percent in absolute value.
4. Synergistic use of FIA plot data and satellite imagery at a national scale likely will benefit both USGS EDC's MRLC 2000 program and the FIA program of the USDA Forest Service.

ACKNOWLEDGMENT

This study was made possible in part by the Raytheon Corporation under U.S. Geological Survey contract 1434-CR-97-CN-40274. The authors want to thank the Fire Science Lab of the Rocky Mountain Research Station of the USDA Forest Service and the Utah GAP Analysis Program of Utah State University for providing the independent test data set.

LITERATURE CITED

Bauer, E.; Kohavi, R. 1998. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Machine Learning. 5: 1-38.

Chan, J.C.-W.; Huang, C.; DeFries, R.S. 2001. Enhanced algorithm performance for land cover classification using bagging and boosting. IEEE Transactions on Geoscience and Remote Sensing. 39(3): 693-695.

Franklin, J.; Logan, T.L.; Woodcock, C.E.; Strahler, A.H. 1986. Coniferous forest classification and inventory using Landsat and digital terrain data. IEEE Transactions on Geoscience and Remote Sensing. GE-24(1): 139-149.

Friedl, M.A.; Brodley, C.E.; Strahler, A.H. 1999. Maximizing land cover classification accuracies produced by decision trees at continental to global scales. IEEE Transactions on Geoscience and Remote Sensing. 37(2): 969-977.

Hall, F.G.; Townshend, J.R.; Engman, E.T. 1995. Status of remote sensing algorithms for estimation of land surface state parameters. Remote Sensing of Environment. 51: 138-156.

Hansen, M.; Dubayah, R.; DeFries, R. 1996. Classification trees: an alternative to traditional land cover classifiers. International Journal of Remote Sensing. 17(5): 1075-1081.

Huang, C.; Davis, L.S.; Townshend, J.R.G. 2002a. An assessment of support vector machines for land cover classification. International Journal of Remote Sensing. in press.

Huang, C.; Wylie, B.; Homer, C.; Yang, L.; Zylstra, G. 2002b. Derivation of a Tasseled cap transformation based on Landsat 7 at-satellite reflectance. International Journal of Remote Sensing. in press.

Irish, R.R. 2000. Landsat 7 science data user's handbook, Report 430-15-01-003-0. National Aeronautics and Space Administration, http://ltpwww.gsfc.nasa.gov/IAS/handbook/handbook_toc.html.

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

Michie, D.; Spiegelhalter, D.J.; Taylor, C.C., eds. 1994. Machine learning, neural and statistical classification. New York, NY: Ellis Horwood. 289 p.

Mickelson, J.G.; Civco, D.L.; Silander, J.A. 1998. Delineating forest canopy species in the northeastern United States using multi-temporal TM imagery. Photogrammetric Engineering & Remote Sensing. 64(9): 891-904.

Quinlan, J.R. 1993. C4.5 programs for machine learning. San Mateo, CA: Morgan Kaufmann Publishers. 302 p.

Townshend, J.R.G. 1992. Land cover. International Journal of Remote Sensing. 13(6): 1319-1328.

Vogelmann, J.E.; Howard, S.M.; Yang, L.; Larson, C.R.; Wylie, B.K.; Driel, N.V. 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 & Remote Sensing. 67(6): 650-662.

Yang, L.; Homer, C.; Hegge, K.; Huang, C.; Wylie, B. 2001a. A Landsat 7 scene selection strategy for a national land cover database. In: IEEE International geoscience and remote sensing symposium; 2001 July 9-13; Sydney, Australia. Piscataway, NJ: Institute of Electrical and Electronics Engineers, Inc. CD ROM, 1 disk.

Yang, L.; Stehman, S.V.; Smith, J.H.; Wickham, J.D. 2001b. Thematic accuracy of MRLC land cover for the eastern United States. Remote Sensing of Environment. 76(3): 418-422.

Zhu, Z.; Yang, L.; Stehman, S.V.; Czaplewski, R.L. 2000. Accuracy assessment for the U.S. Geological Survey regional land cover mapping program: New York and New Jersey region. Photogrammetric Engineering & Remote Sensing. 66(12): 1425-1435.