# APPLICATION OF DECISION-TREE TECHNIQUES TO FOREST GROUP AND BASAL AREA MAPPING USING SATELLITE IMAGERY AND FOREST INVENTORY DATA

**George Xian, Zhiliang Zhu**
Raytheon, USGS EROS Data Center
47914 252nd Street, Sioux Falls, SD 57198
xian@usgs.gov, zhu@usgs.gov

**Michael Hoppus**
US Forest Service
Northeastern Research Station
5 Radnor Coporate Center, 100 Matsonford Rd – Suite 200, Radnor, PA 19087

**Michael Fleming**
Raytheon, EROS Data Center
Alaska Field Office
fleming@usgs.gov

## ABSTRACT

Accurate, current, and cost-effective fire fuel data are required by management and fire science communities for use in reducing wildland fire hazards over large areas. In this paper we present results of applying decision-tree techniques to mapping vegetation parameters (such as vegetation types and canopy structure classification) required for fire fuel characterization. Specifically, we present preliminary results of mapping forest types and average basal area by different forest types at 30-meter resolution. Input data into the decision tree model included Landsat-7 ETM+ spring, summer and fall greenness, brightness and wetness of the tasseled cap transformation, topographic data layers such as slope and elevation, and forest variables measured on inventory plots in the Mid-Atlantic region. Using decision-tree models, eight forest types were successfully identified in training cases and mapped for the entire mapping area. Forest basal area per unit area (conifer and deciduous) was estimated as well using regression-tree models. Cross-validation conducted for both forest types and basal area showed that discrete forest type estimation error was 35% and continuous basal area relative errors were between 58 and 72%. Accuracy was higher in homogeneous forested lands and lower in areas with fragmented forest cover. The study demonstrated that decision tree and regression tree methods are efficient for large-area vegetation mapping if sufficient large-amount of reference data are available.

## INTRODUCTION

Fire fuel data are important for wildland fire suppression planning and hazard assessment. Research in complex fire behavior and fire effects has the need for accurate fire fuel classification and estimation over large area. Mapping fire fuels requires estimation of vegetation structure. Remote sensing data have been used to estimate forest structure for almost two decades. Successful works include modeling canopy optical-geometric properties (Li and Strahler 1985; Li and Strahler 1992). Many researchers have used the modeling method with remote sensing data and classification algorithms to estimate vegetation cover. These include using Landsat TM data and Li and Strahler's Geometric-Optical model to estimate crown cover projection, canopy size, and tree density in Southeast Queensland Australia, with the overall mapping accuracy of 67.4% (Scarth and Phinn 2000); the California multi-attribute vegetation mapping for forest service lands in California using TM data and digital elevation model (Franklin et al. 2000). However, canopy models are computationally intensive, species specific, and often are not inversable. To effectively classify land cover from remotely sensed data, many researches have adopted tree-based techniques such as decision tree and regression tree. Successful effects include Friedl and Brodley (1997), Borak and Strahler (1999), and Friedl et al. (1999). Therefore there is great interest in testing the tree-based techniques for

mapping certain vegetation cover and structure variables for use in large-area fire fuels characterization. In this study, we tested decision tree technique and its derived method – regression tree to map spatial distribution of forest types and forest basal areas. This research was conducted in a mapping area located in the mid-Atlantic region that is also USGS mapping zone 60. Seasonal vegetation data obtained from Landsat 7 images and topographic information were used in the research.

## MAPPING FOREST TYPES USING TREE-CLASSIFICATION

Using a commercial tree-classification package, C5.0 (www.rulequest.com), we applied decision tree classifier (Quinlan 1993; Freund and Schapire 1996; Qu inlan 1996; Friedl et. al. 1999) to mapping forest types defined by the Society of American Foresters (Eyre 1980). As part of this research effort, scientists traveled to Forest Service to extract relevant forests inventory data, merged the data with sampled Landsat 7 tassel cap transformation bands, then removed all geographic coordinates to build up a training data set. A recent study (Huang et. al. 2002) showed that tassel cap transformation from Landsat 7 at satellite reflectance was a useful tool for compressing spectral data into a few bands associated with physical scene characteristics. In our research, we applied the tassel cap spring, summer and fall brightness, greenness, wetness transformed from Landsat-7 ETM+ reflectance band 1-5 and 7. Topographic attributes included slope and digital elevation model. Eight different forest types were classified. The

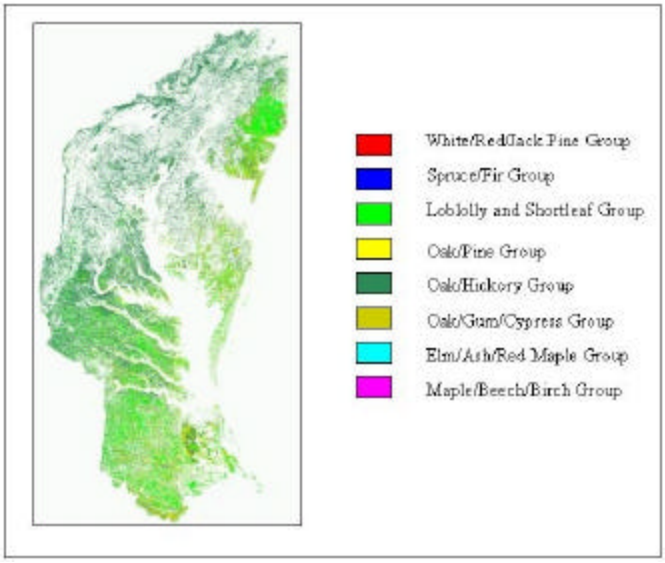


Figure 1 The model forest type classification on the mapping zone

boosting technique (Quinlan 1996; Friedl et. al. 1999) for improving classification accuracy was adopted in the process. To evaluate predictive accuracy of the classifier, we used f-fold cross validation for training cases. All cases were divided into f blocks of the same size and class distribution. A classifier in a block was constructed from the cases in the remaining blocks and tested on the cases in the hold-out block. The error rate of a classifier produced from all the cases was estimated as the ratio of the total number of errors on the hold-out cases from all folds to the total number of cases. This gives an indication of how a decision tree trained from this data would be expected to perform on independent data or the robustness of the model. Huang et. al. (2001) has shown good agreement between cross validation accuracy and independent test data accuracy. The cross-validation result showed that the estimation accuracy was 65.2%, and the standard error was 2.4%. Based on the training cases results, a decision tree model was formulated from a number of decision rules and used to estimate forest type distribution in this region. Figure 1 is the forest type classification on the mapping zone from the decision tree model estimation. The figure shows that Loblolly and Shortleaf, Oak/Pine zone and Oak/Hickory groups cover most area in this region. Other types of tree can be seen on the mapping zone as scattered patches.

## FOREST BASAL AREA ESTIMATION USING REGRESSION TREE TECHNIQUE

Forest inventory data include several forest type basal area values (square feet per acre) for the mapping zone. We display three dimensional plots of coniferous and deciduous basal area values in this region in Figure 2 and Figure 3, respectively. The values of coniferous basal area (CBA) are high in both northeast and southwest sides. Deciduous basal area (DBA) has two apparent large value peaks in both central south and north regions.

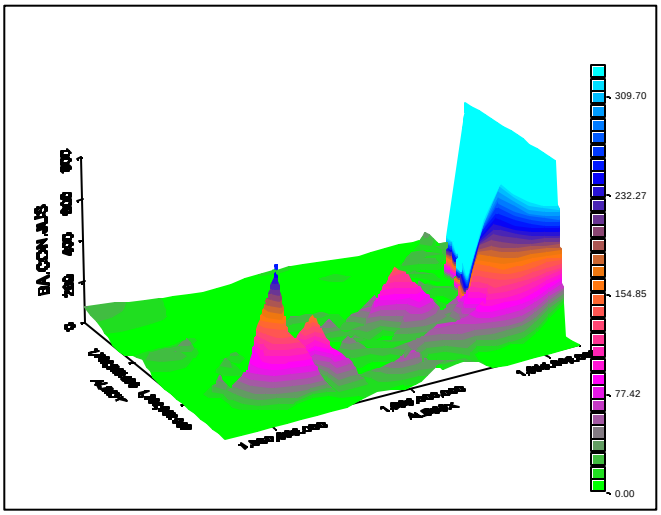


Figure 2 3-D surface plot of Forest Inventory database: Conifer basal area

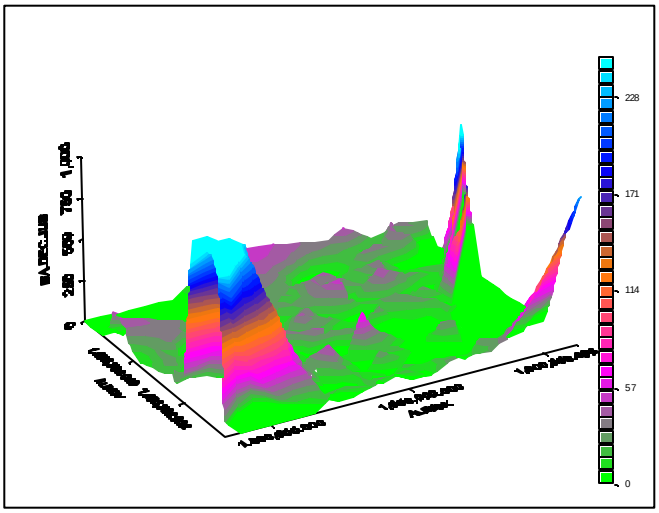


Figure 3 3-D surface plot of Forest Service Inventory database: Deciduous basal area

To predict continuous forestry basal area using remote sensing data, we chose a modified regression-tree-based predictive model, Cubist (www.rulequest.com), that is a sister system of decision-tree classification model. The modified regression-tree-based model estimates a case's target value in terms of its attribute values by building a model containing one or more rules, where each rule is a conjunction of conditions associated with a linear expression. A rule indicates that, whenever a case satisfies all the conditions, the linear model is appropriate for predicting the value of the target attribute. The rule-based predictive model thus resembles a piecewise linear model (except that the rules can overlap) based on training data.

In this study, we set out to predict values of deciduous and conifer basal areas. Training data included onsite deciduous and conifer basal area measurements from Forest Service inventory database; Landsat 7 tasseled cap greenness, wetness and brightness; slope and digital elevation model data. To improve the model prediction accuracy, booting was also adopted for constructing more accurate regression model.

## Deciduous tree basal area prediction

The regression tree rules were developed for target values based on attributes from training data set. A total of twenty-seven rules were generated. These rules included almost all attributes listed in the training data. After rules of the target value were built for all conditions, the linear model was applied to predicting values of target attributes based on each rule conditions and Landsat 7 ETM data associated with geographic slope and elevation over the whole mapping region. All target values were estimated for 30 meter resolution pixels in the region. The model prediction of deciduous forest basal area over the region was plotted in Figure 4. Most large values of basal area were located around the central and north parts of the mapping zone where the land slope and altitude value are relatively high.

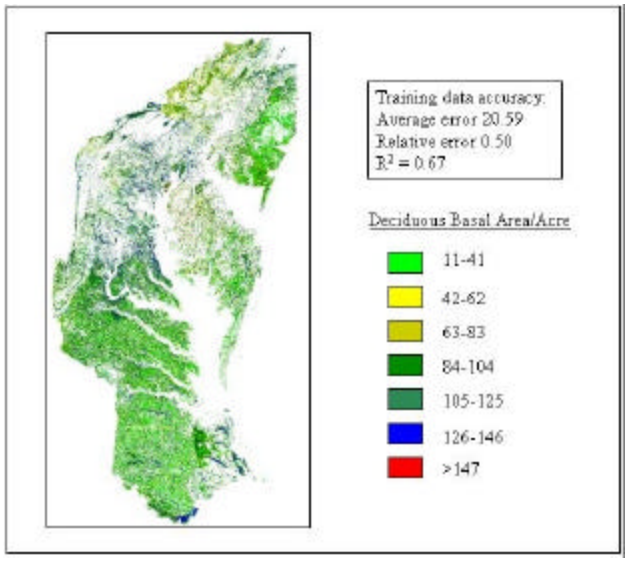


Figure 4 Model estimation of Deciduous Basal Area

## Coniferous basal area prediction

From the Forest Service inventory data plot in Figure 2 we know that coniferous forest basal area has relative large values in the northeast and southwest parts of the mapping zone. To estimate the magnitude of CBA over the region, we used the same regression tree model and data set used for deciduous trees. The only change in the model was that the target attribute became CBA. Running from training data set, a rule-based predictive model was built up with twenty-six rules for the target value. Then, the model was applied to the entire region using prepared Landsat 7 ETM+ images. The model prediction of CBA over the region was displayed (Figure 5). Large values in both southwest and northeast regions were very apparent. Generally, the linear regression model prediction showed that the distribution of modeled basal area values were very close to Forest Service Inventory data shown in Figure 2. The lack of coniferous forests in northwest region of the mapping zone is apparent. In the southern part of the mapping zone, coniferous and deciduous forests mix with relative large basal areas for both trees.

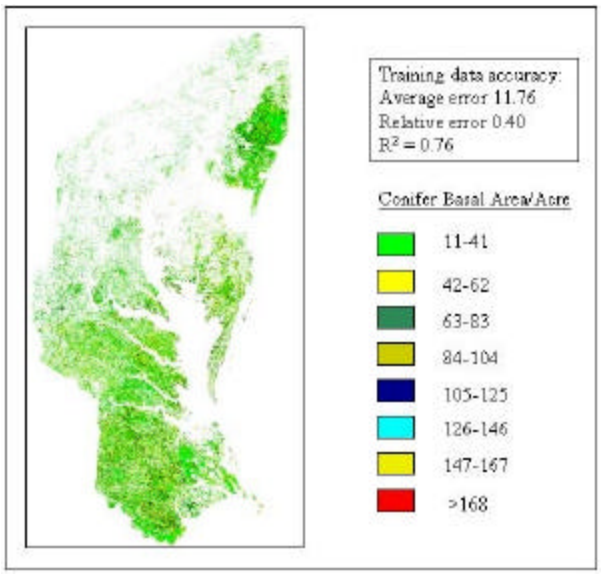


Figure 5 Model estimation of Coniferous Basal Area

## Validation of model prediction

The accuracy of the model prediction needs to be assessed. The prediction of the regression tree model on the training cases from which it was constructed was evaluated by calculating the regression tree model average error, relative error, which is the ration of the average error to the error that would result from always predicting the meaning value, and $R^2$. These estimations were obtained from training examples. We needed to validate the predictive continuous variables of the model on new cases. Similar with the decision-tree classification cross validation to assess predictive accuracy, we used f-fold cross-validation here to obtain a reliable estimate of predictive accuracy. The existing training cases were divided into f blocks of nearly the same size and target value distribution. The accuracy of a model produced from all the cases was estimated by averaging results on the hold-out cases.

The accuracies of regression tree model calculation and prediction were shown in Table 1.

Good model performance requires the relative error value to be less than 1. From our training cases, the model relative error was 0.40 and 0.50 for conifer and deciduous forest basal areas, respectively. The $R^2$ reached 0.76 for conifer and 0.67 for deciduous. The result suggested that the model had relative higher consistency for conifer than for deciduous over the whole mapping region. The cross validation results showed that $R^2$ equals to 0.46 for conifer forest and 0.38 for deciduous forest.

Table 1. Accuracies of regression-tree model prediction

| Items | Tree model relative error | Tree model $R^2$ | Cross validation relative error | Cross validation $R^2$ |
|---|---|---|---|---|
| Conifer Basal Area | 0.40 | 0.76 | 0.58 | 0.46 |
| Deciduous Basal Area | 0.50 | 0.67 | 0.72 | 0.38 |

## Comparison between K-NN and regression-tree model

The k-nearst neighbor (K-NN) classifier is a simple classification method that has been found effective for mapping forest structure variables (Michie, 1994; Tomppo, 1996). Generally, K-NN method estimates a field parameter at a particular location by a weighted average of k nearest neighbor (known) pixels. The weighted function is distance based. The number of nearest neighbors, k, is selected such that the k nearest distances relative to the pixel location are within this chosen range. One of authors had applied this technique to a small area in the

central east region of the mapping zone. The forest basal areas were estimated using K-NN and Forest Inventory database. Here, we compared the values of model predictions for both conifer and deciduous forest basal areas from regression tree and K-NN models with Forest Inventory field data. A linear model fit generated the linear residual standard error and $R^2$ values from two technique predictions associated with real values.

Table 2. Comparison between two model prediction and Forest Inventory data

| | Residual Standard Error | | $R^2$ | |
|---|---|---|---|---|
| Estimation | K-NN | Tree-Model | K-NN | Tree-Model |
| Conifer Basal Area | 25.29 | 24.23 | 0.2622 | 0.4969 |
| Deciduous Basal Area | 23.97 | 18.63 | 0.6124 | 0.7659 |

Table 2 presents comparison of results between regression tree and KNN model predictions and observed values using linear regression fit. The regression tree model had smaller residual standard error and large $R^2$ values than K-NN for the same forest type. The regression tree model predictions had less standard errors than KNN. Figure 6 and 7 are linear regression fit plots for conifer and deciduous forests predictions, respectively. We also noticed that deciduous forest basal area predictions in KNN and regression tree models had standard errors of 0.0984 and 0.0846, respectively. The conifer forest basal area predictions had standard errors of 0.1233 and 0.1181 from KNN and tree model, respectively. The regression tree model had larger $R^2$ values and less standard errors than K-NN model. Deciduous basal area prediction was closer to the observed measurements in this area. This result was different from the cross examination we discussed in section 3.3 where deciduous prediction had less accuracy than conifer. This difference might be caused by the difference in the locations where the models were applied. The comparison data were from a small region where deciduous forest was majority vegetation cover. The model prediction might be in favor to it in this small area. But for the entire mapping region, the average result could be different.

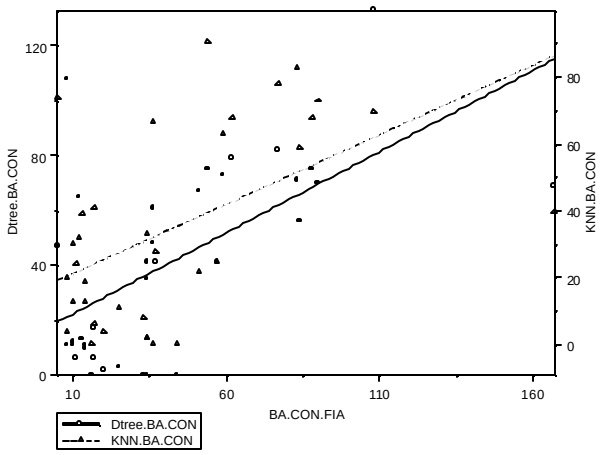


Figure 6 Comparison of conifer basal area prediction from Regression-Tree and K-NN models to the Forest Inventory data

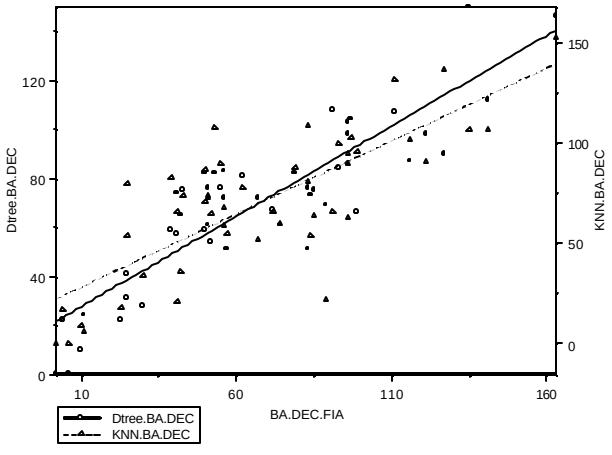


Figure 7 Comparison of deciduous forest basal area prediction from Regression-Tree and K-NN models to the Forest Inventory data

## CONCLUSIONS

Forest types and average basal area, required for fire fuels characterization, were mapped in USGS mapping zone 60. The decision tree approach was used in this study. The training data included Landsat 7 data and Forest Inventory database. Eight different forest types were classified in the region. The Loblolly and Shortleaf Group forests were apparent in the east and southeast part of the region. Oak/Hickory Group covers most of north and northwest region. The predictive accuracy of decision tree model from training data set reached 65.2%. By using the same Landsat imagery and Forest Service Inventory database basal area on site measurement data, we used the regression tree model to estimate the deciduous and conifer forests basal area values over the whole mapping region. The regression tree model prediction showed that most of large basal area values (>100) of deciduous forest were in the northwest and northeast of the region. The result was consistent with the model classification conclusion and Forest Service Inventory data. The conifer forest basal area estimation showed that most of large values were in southern part of the region. In general, the conifer forest basal area prediction was more accurate than the deciduous forest results.

We also compared regression tree model results with results obtained from K-NN model in a small area in central-east part of the mapping region. Two model predictions were compared with Forest Service inventory data using linear regression fit. The regression tree model results were closer to the Forest Service Inventory data.

## ACKNOWLEDGMENT

The authors would like to thank US Forest Service for supplying Forest Service Inventory database. This study was made possible in part by the Raytheon Corporation under US Geological Survey contract 1434-CR-97-CN-40274.

## REFERENCES

Eyre, F.H. (Ed) (1980). Forest Cover Types of the United States and Canada. Society of American Foresters.

Franklin, J., C.E. Woodcock, and R. Warbington (2000). Multi_attribute Vegetation Maps of Forest Service Lands in California Supporting Resource Management Decisions. Photogrammetric Engineering & Remote Sensing **66**(10): 1209-1217.

Freund, Y. and R.E. Schapire (1996). A decision-theortic generalization of on-line learning and an application to boosting. 2nd Euro. Conf. EuroCOLT96

Friedl, M. A., C.E. Brodley, A. H. Strahler (1999). Maximizing Land Cover Classification Accuracies Produced by Decision Trees at Continental to Global Scales. *IEEE Transactions on Geoscience Remote Sensing* **37**(2): 969-977.

Huang, C., L. Yang, C. Homer, M. Coan, R. Rykus, Z. Zhang, B. Wylie, K. Hegge, Z. Zhu, A. Lister, M. Hoppus, R. Tymcio, L. DeBlander, W. Cooke, R. McRoberts, D. Wendt, and D. Weyerman (2001). Synergistic Use of Data and Landsat 7 ETM+ Images For Large Area Forest Mapping. Thirty Annual Midwest Mensurationists Meeting and Third Annual Forest Inventory and Analysis Symposium. Traverse City, MI., USA.

Huang, C., B. Wylie, L. Yang, C. Homer, G. Zylstra (2002). Derivation of a tasselled cap transformation based on Landsat 7 at satellite reflectance. *International Journal of Remote Sensing*

Li, X., and A. H. Strahler (1985). Geometric-optical modeling of a conifer forest canopy. *IEEE Transactions on Geoscience Remote Sensing*. GE-23, 705-721.

Li, X., and A. H. Strahler (1992). Geometrical-optical bidirectional reflectance modeling of the discrete-crown vegetation canopy: Effect of crown shape and mutual shadowing. *IEEE Transactions on Geoscience Remote Sensing* **GE-30**: 276-292.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers.

Quinlan, J. R. (1996). Bagging, boosting, and c4.5. 13th National Conference of Artificial Intellegence, Portland, OR.

Michie, D., D.J. Spiegelhalter, and C.C. Taylor (editors) (1994). Machine Learning, Neural and Statistical Classification. Ellis Horwood, Hertfordshire, UK, 289p.

Scarth, P., S. Phinn (2000). Determining Forest Structural Attributes Using an Inverted Geometric-Optical Model in Mixed Eucalypt Forests, Southeast Queensland, Australia. *Remote Sensing Environment.* **71**: 141-157.

Tomppo, E. (1996). Multi-source National Forest Inventory of Finland: New Thrusts in Forest Inventory. Proceedings to the New Zealand Preharvest Inventory. Scandinavia Journal of Forest Research, 14, 182-192.