USDA United States Department of Agriculture

Forest Service

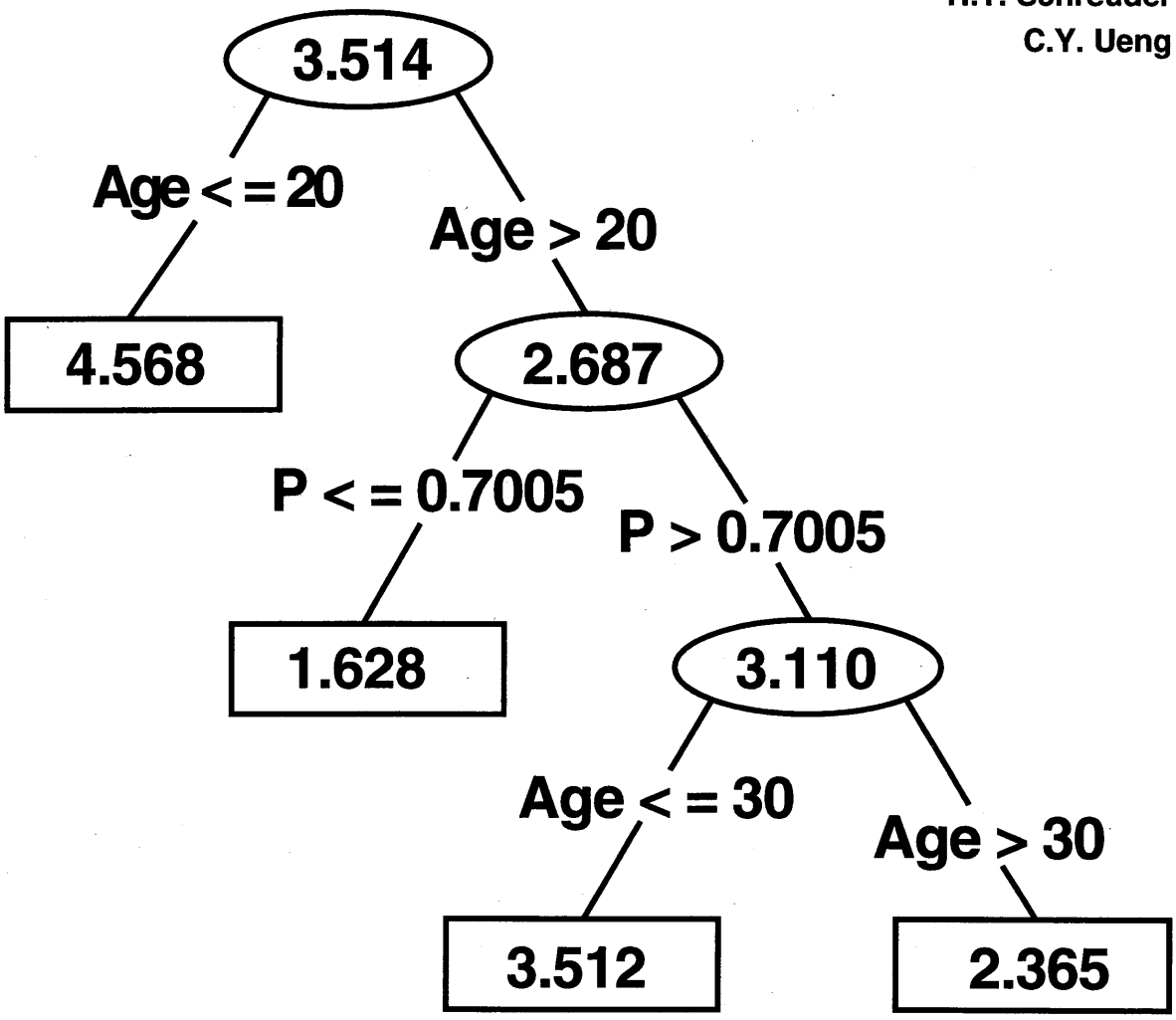
Rocky Mountain Research Station

Fort Collins, Colorado 80526

Research Paper RMRS-RP-2

A Nonparametric Analysis of Plot Basal Area Growth Using Tree Based Models

G.L. Gadbury
H.K. Iyer
H.T. Schreuder
C.Y. Ueng



Abstract

Gadbury, G.L.; Iyer, H.K.; Schreuder, H.T.; and Ueng, C.Y. **A nonparametric analysis of plot basal area growth using tree based models.** Res. Pap. RMRS-RP-2. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 14 p.

Tree based statistical models can be used to investigate data structure and predict future observations. We used nonparametric and nonlinear models to reexamine the data sets on tree growth used by Bechtold et al. (1991) and Ruark et al. (1991). The growth data were collected by Forest Inventory and Analysis (FIA) teams from 1962 to 1972 (4th cycle) and 1972 to 1982 (5th cycle). We used tree based models to group observations into clusters that were specified by covariate values. Next, we performed a permutation test on the grouped data to test for a change in tree growth rates from the 4th cycle to the 5th cycle. Our techniques differed from those used by Bechtold et al. (1991) and Ruark et al. (1991). The data was not assumed to follow any parametric distribution, the relation between response and covariates was not assumed to be linear, and the test for a change in tree growth did not require any parametric assumptions. The methodology presented here is general and applicable to other situations where the significance of a specific covariate is in question. Despite these relaxed constraints of analysis, our results generally agreed with those of Bechtold et al. and Ruark et al.

Key words: nonparametric, nonlinear, models, tree based statistical models, tree growth rates, parametric distribution, tree growth data

The Authors

G.L. Gadbury is a statistics student at Colorado State University.

H.K. Iyer is a statistics professor at Colorado State University.

H.T. Schreuder is a mathematical statistician at the USDA Forest Service, Rocky Mountain Station in Fort Collins.

C.Y. Ueng is a Ph.D. graduate from the Department of Statistics at Colorado State University.

Publisher

Rocky Mountain Research Station
Fort Collins, Colorado
February 1998

You may order additional copies of this publication by sending your mailing information in label form through one of the following media. Please send the publication title and number.

Telephone	(970) 498-1719
E-mail	rschneider/rmrs@fs.fed.us
DG message	R.Schneider:S28A
FAX	(970) 498-1660
Mailing Address	Publications Distribution Rocky Mountain Research Station 3825 E. Mulberry Street Fort Collins, CO 80524-8597

Cover: Tree based model.

A Nonparametric Analysis of Plot Basal Area Growth Using Tree Based Models

G.L. Gadbury, H.K. Iyer, H.T. Schreuder, and C.Y. Ueng

Contents

Introduction

Since 1928, the Forest Inventory and Analysis (FIA) units of the USDA Forest Service have surveyed the forest resources of the United States. Estimates of aggregates, for example, area in major land classes and/or forest types, changes in areas and volumes over time, and changes in the forest resource over large areas, are identified. In recent decades, these inventories have produced data that, upon subsequent analyses, have revealed a decrease in growth of natural southern pine stands in the southeastern United States from 1972 to 1982 relative to 1962 to 1972 (Sheffield et al. 1985, Sheffield and Cost 1987, Zahner et al. 1989). This reported decrease has been as high as 23% for loblolly pine (*Pinus taeda* L.), 32% for Georgia shortleaf pine (*P. echinata* Mill.), 28% for Georgia slash (*P.elliotti* Engel.), and 29% for Alabama longleaf pine (*P. palustris* Mill.) (Bechtold et al. 1991, Ruark et al. 1991, hereafter Bechtold et al. and Ruark et al.).

These findings may cause concern because questions about the potential reasons for the reported decline in growth become relevant, and such data do not establish cause-effect (Schreuder and Thomas 1991). For example, newspapers have discussed the possibility of a pollution effect even though such a link has not been established.

Considering the seriousness of the implications of a growth decline reaching as high as 32%, an analysis of the 2 data sets used by Bechtold et al. and Ruark et al. is necessary. Alternative analyses of a given data set are reasonable because statistical models are approximations of reality, at best, and no single analysis can consider all peculiarities associated with any given data set. Although the models used by Bechtold et al. and Ruark et al. have been subjected to careful scrutiny by others, some issues remain unexplored. Are there other models, perhaps nonlinear ones, that explain the data as well as or better than the Bechtold et al. or Ruark et al. models? Do these alternative models yield results that are consistent with earlier findings? Could these data be analyzed using statistical methods requiring fewer assumptions?

In this paper we consider a relatively recent regression technique, referred to as either tree based regression, classification and regression trees (CART; Breiman et al. 1984), or tree based models (Clark and Pregibon 1993). The models are so named because of their analogies to actual trees. For instance, developing a model is called "growing" the tree, refining a model is called "pruning" the tree, and tree models have branches and leaves.

This regression technique is useful when complex unknown relationships exist among large numbers of variables. Because the procedure is not based on linear modeling or normal distributions, we felt that it would lead to a nonredundant analysis of the data sets. Such a procedure would be useful to assess other changes of interest in FIA and national forest inventory data.

In this paper, we review the literature, outline the principles underlying the tree based model procedure, and present an approach for assessing the difference in tree growth from the 4th to 5th cycles involving a nonparametric test. This approach is general and is applicable to many different problems where the significance of a specific variable is in question. We report and discuss the results of our analyses and describe a simulation study that gives some insight about how results are affected by model selection and the correlation structure of the data.

Literature Review

Bechtold et al. classified the inventory plots in Georgia as loblolly, shortleaf, and slash pine stands. For each type of plot, they analyzed data representing "all" trees and data representing "pine only" trees so a total of 6 analyses were conducted. The growth rates of loblolly and slash pine plots were investigated using a model of the form:

$$ln(G)=b_0*c_1+b_1*c_2+b_2*S+b_3*ln(A)$$

$$+b_4*ln(N)+b_5*P+b_6*ln(M+1)+\epsilon \qquad (1)$$

where c_1 and c_2 are indicators for cycle 4 and cycle 5, and other variables are defined in table 1.

Model 1 was used to analyze "pine only" and "all trees" data. An exception to the model given by equation 1 is where Bechtold et al. used an interaction term involving number of stems per acre (N) and cycle for shortleaf pine,

Table 1. Variable descriptions for Model 1.

Variable	Description
G	gross annual basal area growth per acre (survivor growth + ingrowth)
S	site index representing volume growth potential; a relation between age and height of dominant and co-dominant pines in each stand (base 50 years)
A	stand age (midpoint of 10 year class)
N	number of stems per acre
P	ratio of yellow pine basal area per acre to basal area of all species
M	annual basal area mortality per acre of trees ≥1 in. dbh (diameter breast height) alive at initial inventory that die from natural causes prior to terminal inventory
ϵ_1	error terms assumed *iid Normal* $(0,\sigma^2)$

instead of a common N term across cycles as in Model 1. All cases showed a significant reduction in growth from cycle 4 to cycle 5 ranging from 16% for loblolly "all trees" to 32% for shortleaf "all trees."

Ruark et al. used a different data set for their analysis and used the model:

$$ln(PSGi)=b_0*c_1+b_1*c_2+b_2*S+b_3*ln(QMD)$$

$$+b_4*ln(N)+b_5*P+b_6*ln(M+1)+\epsilon_2 \qquad (2)$$

where $PSG1$ is pine survivor growth of trees ≥1.0 in. dbh (diameter at breast height), $PSG5$ is pine survivor growth of trees ≥5.0 in. dbh, initial QMD (quadratic mean diameter) of trees ≥1.0 in. dbh; other variables are defined in table 1. Ruark et al. performed 2 analyses using Model 2, one for trees ≥1.0 in. dbh and the other for trees ≥5.0 in. dbh (a subset of the former). They analyzed tree growth for Georgia and Alabama and, with the exception of Alabama loblolly $PSG5$ growth, all results showed a significant reduction on $PSG1$ and $PSG5$ growth from cycle 4 to cycle 5 ranging from 10% to 31%.

Ouyang et al. (1992) suggested that the analyses used by Bechtold et al. and Ruark et al. could be improved using a procedure based on bootstrap and weighted jackknife confidence intervals, still fitting the same linear model at each iteration. With their procedures, they confirmed the growth reductions found by Bechtold et al. and Ruark et al.

The data sets of Bechtold et al. and Ruark et al. were screened to obtain similar stands of timber for each period. Refer to Bechtold et al. and Ruark et al. for their respective screening criteria. Although the criteria used to select plots were the same for both periods, the samples taken in the 2 periods may not represent the same population; therefore, the population of inference remains unknown. For example, if there was visual evidence of disturbance on a plot, the plot was screened out. Disturbance rates can be high for southern pine stands.

Models 1 and 2 were selected for reasons relating to biological sensibility in addition to obtaining a reasonable statistical fit. Possibly, adding other terms, such as 2nd order terms (quadratic and two-way interactions), might lead to a substantially better fit. Bechtold et al. and Ruark et al. found significant interactions among covariates and between covariates and cycle, although inclusion of these terms in their model only slightly affected predictions. Different approaches for model selection (all subsets regression, backward elimination, stepwise regression, etc.) will generally lead to models different from Models 1 and 2. Such models will probably fit the data adequately and may be as biologically meaningful. Another factor that may influence the results of a model selection exercise is the initial "pool" of covariates; the covariates themselves, their powers, products, etc. Generally, there is a large degree of uncertainty associated with the model selection exercise. In this paper, we do not deal with uncertainty in model selection but do consider an entirely different class of models that offer advantages when the data structure is complex. We describe these models in the next section.

Tree Based Models

Tree based models can be used to study either classification or regression problems. Although we are concerned with the latter, the 2 problems are closely related; the regression problem is solved by classifying continuous observations into homogeneous subsets. The technique was proposed by Morgan and Sonquist (1963) as a means of analyzing large scale survey data, and it found application in the medical community where it was used to classify patients into high or low risk categories. The technique has also been used in a variety of other classification and regression problems (Verbyla 1987, Ciampi et al. 1988, Temkin et al. 1995), Chaudhuri et al. 1995).

As indicated, the technique is often referred to as classification and regression trees (CART) (Breiman et al. 1984). In forestry, Le May (1994) recommends the use of an analysis, such as CART, to develop objective rules to classify whether standing trees are sound or decayed. In this paper, we use the statistical package S-plus that contains many useful functions for developing and analyzing

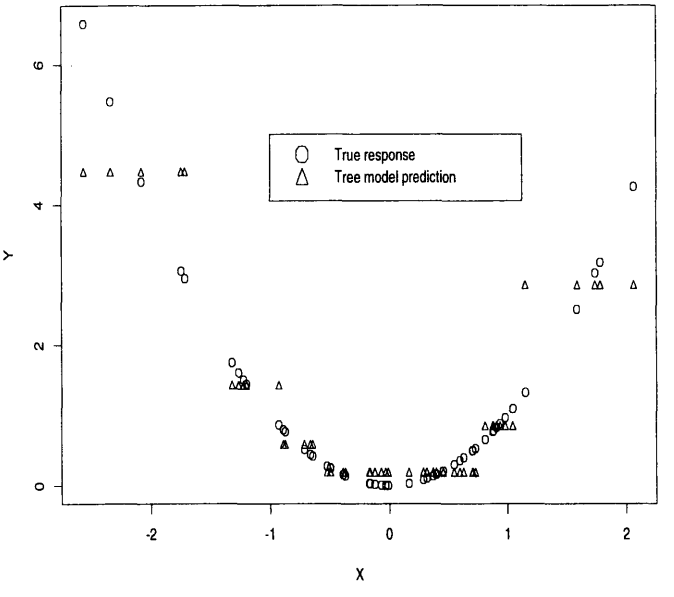


Figure 1. The fit of a tree based model where X is 50 observations from a standard normal distribution, and $Y=X^2$. The tree model predictions approximate the quadratic relationship with a step function.

tree based models (see Clark and Pregibon 1993 for a treatise of tree based models in S-plus).

A tree based model attempts to fit a step function to the response variables, where the steps are determined by covariate values. The relationship between the response variable and the covariates can be simple or complex. It is not required that the data follow any particular parametric distribution. Figure 1 illustrates the basic idea underlying tree based models. In this figure, X is a vector of 50 observations from a standard normal population and, for illustrative purposes, we assume there is a perfect quadratic relation between the response and covariates, $Y=X^2$. The tree based model approximates this quadratic relation with a step function. Although the tree model does not discover the exact relationship between Y and X, it determines the existence of curvature without starting from an assumed model form such as $Y=a+bX+cX^2$, as would a classical regression analysis. For the data displayed in figure 1, we can examine the plot first and then decide to fit a model of the form $Y=a+bX+cX^2$. But for the FIA data, visualization is difficult because of high dimensionality. No model form may be postulated *a priori* with any degree of conviction. We believe that the data structure in the aforementioned data sets is sufficiently complex to justify the tree based model approach.

Tree based models have certain additional attractive features. They are invariant to monotonic transformations of predictor variables, and they automatically attempt to accommodate higher order terms and interactions among covariates. However, they are not invariant to transformations of response since prediction errors are determined using the scale for the response. Outliers affect a tree model differently from a linear model. A tree model attempts to cluster responses into homogeneous groups, and an outlier only affects one of these groups no matter how extreme the outlier is. In contrast, an extreme outlier in a linear model may influence the entire model.

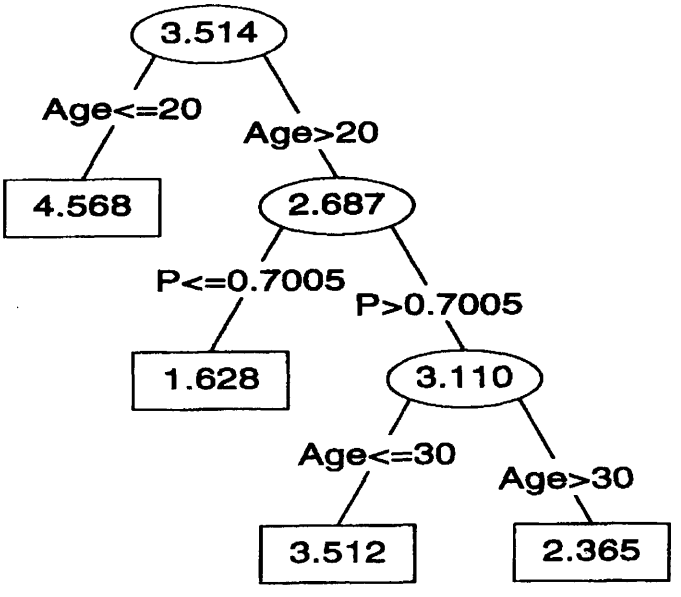


Figure 2. Tree model showing 4 terminal nodes or leaves represented by rectangles. Ovals are intermediate nodes. Numbers inside the ovals and rectangles are the predictions at that node. The top oval node with the prediction of 3.514 is sometimes called the root or parent node. The 2 nodes with predictions 4.468 and 2.687 are sometimes called children nodes etc.

Let us take a simple example of a tree model and illustrate how it is used for making predictions. Consider the prediction tree in figure 2. This tree is a graphical description for the prediction rule suggested by a tree model fit to hypothetical tree growth data with $Y=growth\ rate$ as the response variable and $X_1=Age$ and $X_2=P=proportion\ of\ yellow\ pine$ as covariates. This tree has 4 leaves or terminal nodes. The node labeled 3.514 is the root node, and the value 3.514 represents the mean Y value of all stands in the data. The root node is split into 2 nodes, one corresponding to the branch $Age\leq 20$ and the other to the branch $Age>20$. The mean Y value for all stands in the data with $Age\leq 20$ is 4.568. Likewise, the mean Y value for all the data with $Age>20$ is 2.687. The process is continued until it is determined that no further branching is necessary.

To predict the Y value for a stand with $Age=25$ and $P=0.6$ we begin with the root node and proceed along the branch labeled "$Age>20$" since $X_1=25$ for this stand. Since $P=0.6$, we take the branch "$P\leq 0.7005$" and reach the leaf labeled 1.628. The predicted Y value for this stand is 1.628. Note that each "leaf" corresponds to a data subset that has very similar Y values. In essence, the data cases have been grouped into 4 clusters and the cluster mean Y value is used to predict the Y value of any new case that is classified as belonging to the cluster in question.

Next, we give a brief explanation of the basic procedure used to fit tree based models to data and to assess the goodness of fit. Certain similarities between linear regression modeling and tree based modeling will be noted. The modeling exercise is divided into 3 stages: 1) developing or "growing" a tree model, 2) refining or "pruning" a tree model, and 3) evaluating the goodness of fit. These 3 stages also occur in standard regression modeling where the first stage is developing an initial regression model, the second stage is examining submodels, and the final stage is comparing the performance of various candidate models.

Growing the Tree Model

Tree based regression seeks to partition observations into homogeneous classes. Classic multivariate cluster analysis seeks to group multivariate observations into classes that are "close" together where closeness is determined by a chosen metric distance (Mardia et al. 1979).

Tree based models differ in that the chosen distance of interest is the squared distance of the Y values (response) within clusters from their mean in that cluster.

We first consider a single response variable and a single covariate. Let Y denote a vector of observations and X a vector of covariate values. Tree structured regression is begun by partitioning the data Y into 2 groups, Y_1 and Y_2. The observations that make up Y_1 and Y_2 are determined by the corresponding values in X. Denote the mean of Y_1 as $\bar{y}_1$, the mean of Y_2 as $\bar{y}_2$, and the overall mean of Y as $\bar{y}$. A value in X, say x_0, is chosen so that values in $X>x_0$ designate observations in Y_1 and values in $X<x_0$ designate the observations in Y_2. This value of x_0 then maximizes the reduction in sum of squares:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2-\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 \qquad (3)$$

where i denotes the 2 subsets of Y, n_i is the number of observations in set i, and n is the number of observations in Y. The observations in Y_1 make up a node of the tree as do the observations in Y_2. The observations in the original Y are in the root node. As nodes are formed off the root node, they are recursively split into subsets in the same manner to maximize the quantity in equation (3). Now, we can generalize the scenario described above by allowing multiple covariates so that at each node not only the x_0 value is sought, but also the covariate is sought that, when split on, maximizes the quantity in equation (3).

A node will not be split further if the observations in that node are determined to be homogeneous or if there are too few to justify further splits (≤5 by default). When this condition is reached, the node is called a "terminal node" or "leaf." The tree growing algorithm terminates when there are no nodes remaining that need to be further split.

There is no statistical test that monitors how far a tree model should grow. Thus, the algorithm tends to overfit the data (similar to overfitting a linear regression model). The "overfit" model is sometimes called a "fully grown tree model." A technique called "pruning" is used to simplify the tree model without sacrificing the goodness of fit. Note the similarity between pruning to remove unnecessary nodes from a tree model and backward stepwise regression to remove unnecessary variables from a regression model.

Pruning a Tree Model

Consider a tree T and an arbitrary subtree T' with the same root node as T. Clark and Pregibon (1993) define a quantity, $D_\delta(T')$, which we call the adjusted sum of squares for the subtree T', by

$$D_\delta(T')=D(T')+\delta size(T') \qquad (4)$$

where δ is a penalty factor called a cost-complexity parameter, $size(T')$ is the number of leaves in tree T', and $D(T')$ is the pooled "within" sum of squares for Y (often called

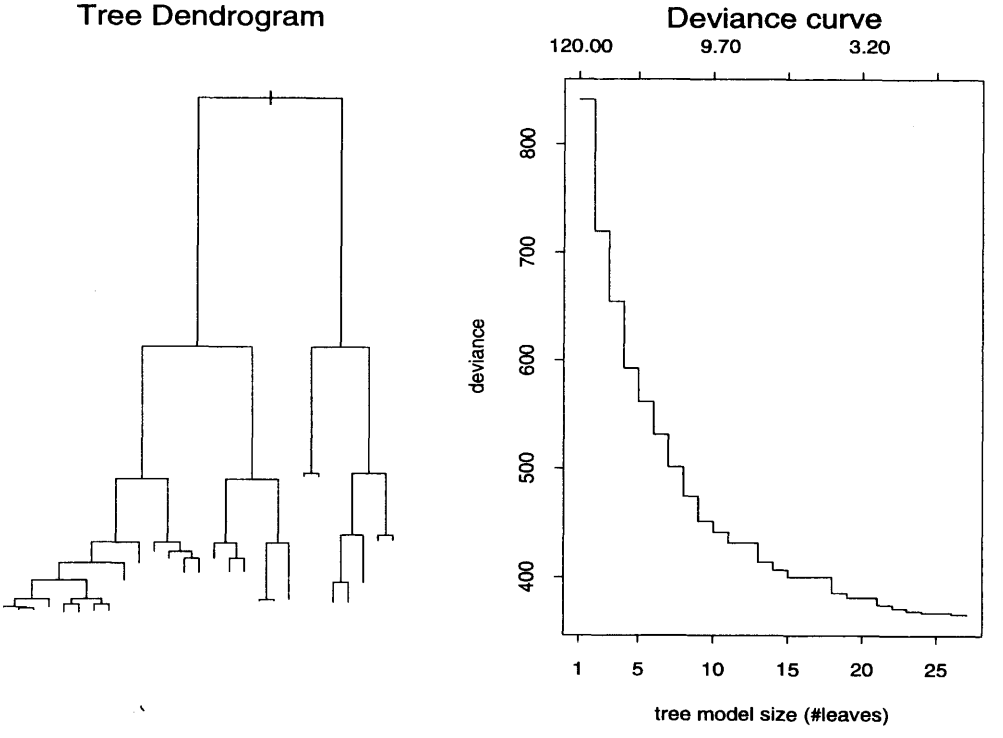


Figure 3. Tree dendrogram with corresponding deviance curve. The length of the vertical lines on the dendrogram is a visual representation of the amount that the tree deviance is reduced by splitting at a particular node. Very short vertical lines at a split indicate that that node will probably be among the first pruned. The values across the top of the deviance curve are δ values that influence tree size.

the deviance), pooled over all the clusters determined by the terminal nodes.

$$D(T') = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \qquad (5)$$

Here k signifies the number of terminal nodes of T', and the other terms are as in equation (3). $D_\delta(T')$ in equation (4) is the adjusted sum of squares for tree T' after incurring a penalty factor related to its size. (This quantity is analogous to Mallows' C_p in the linear regression framework.) The fully grown tree model is pruned, for a given value of δ, to minimize $D_\delta(T')$. So, for δ=0, no pruning is done, and for any positive δ, the terminal nodes that reduce the total tree deviance the least are identified and pruned until $D_\delta(T')$ reaches a minimum for that value of δ. The larger the δ value, the more pruning is required. We can let δ be a vector of values, which creates a sequence of tree sizes and corresponding deviance. A plot of these tree sizes and deviances is a deviance curve. Figure 3 is a dendrogram of a full tree and its corresponding deviance curve.

Goodness of Fit and Predictive Performance

Once a tree model is constructed and pruned for some δ value, one measure of its goodness of fit is its deviance, given by equation (5). The predictive capability of a tree model can be evaluated using cross-validation. If Y is our vector of observations of length n, tree T' is constructed by pruning a fully grown tree T for a chosen value of δ. A new set of observations Y_n and their corresponding X_n covariates are passed through the tree and the resulting prediction errors are calculated. If no new data are available, divide Y into m subsets. Delete the *ith* subset from Y and produce a tree model T'' from the remaining (i-1) subsets of Y, which are pruned to a fixed value of δ. Pass the observations from the *ith* subset through this tree and calculate a total prediction error using the expression:

$$\sum_{j=1}^{k} \sum_{l=1}^{n'_{ij}} (y_{ijl} - \bar{y}_{j(i)})^2 \qquad (6)$$

where l indexes the observations of Y from subset i that fall into node j, $\bar{y}_{j(i)}$ is the mean Y value at node j of tree T'' that was constructed without observations in subset i, n'_{ij} is the number of observations from subset i that fall into node j, and k is the number of terminal nodes in tree T''.

Next, a different subset from 1,2, ... , m is deleted from Y, a new tree model is constructed without this subset, and the calculation of equation (6) is repeated. This is done m times and the m results from equation (6) are summed. The result is the cross validation deviance of the tree model for that particular penalty level of δ. This process is an m-fold cross validation, and the cross validation deviance is a useful indicator of the predictive performance of the tree model. Repeating this process for differing δ values, one can obtain a tree model size (number of terminal nodes) for which predictive performance is optimized for a particular partitioning of observations Y.

Interactions in Tree Based Models

Tree based models automatically attempt to accommodate higher order terms and interactions among the covariates, thus making their use particularly appropriate when the structure of the data is complex. For example, suppose a root node splits on the covariate "Age" to create 2 branches and 2 children nodes of the root node. If one of these children nodes splits on another covariate, "P", to form 2 more branches and 2 grandchildren nodes of the root node, then a form of an "Age:P" interaction results. Figure 2 shows this process. If the data structure is simple, then tree based models are less advantageous. If a simple linear relation among covariates satisfactorily explains the response Y, a tree based model is probably inappropriate.

Methods

We used 2 data sets in this study, the first was used by Bechtold et al., hereafter referred to as Data #1, and the second was used by Ruark et al., hereafter referred to as Data #2.

Data #1 contains information for 3 different pine stands (loblolly, shortleaf, and slash) and for each type there are data for "all" trees and "pine only" trees. In addition to analyzing gross growth for this set of data, we also analyzed net growth. So 12 analyses were conducted on Data #1. The model used for these analyses is:

$$GG = f(A, N, S, P)$$

where GG indicates gross growth and all other variables are described in table 1.

The model for net growth is then:

$$NG = g(A, N, S, P)$$

where NG is net growth and all other variables are described in table 1. Net growth is gross growth minus mortality.

Data #2 comprises measurements of trees from 4 types of pine stands in Georgia (loblolly, longleaf, shortleaf, and slash pine) and 3 in Alabama (loblolly, longleaf, and

shortleaf). For each stand type, information on pine trees ≥1.0 in. dbh and ≥5.0 in dbh was available; the latter is the merchantable or saleable component per Ruark et al. For each stand type, gross and net growth were modeled for trees ≥1.0 in. dbh; a total of 14 analyses. However, only gross growth was modeled for trees ≥5.0 in dbh, a total of 7 analyses, because information on mortality for this subset was unavailable. Growth for Data #2 was modeled as indicated below:

$$PSGi=h(QMD, N, S, P)$$

where the variables are described in the introduction, and i=1,5.

Net growth was modeled by the same equation but with the response replaced by *PSG*1 - *Mortality*.

In the above models none of the variables are transformed and mortality appears (indirectly) only on the left hand side of the net growth model. As mentioned, transformations of covariates are unnecessary since tree based models are invariant to monotonic transformations of predictor variables, and there was no advantage in transforming the response. Secondly, Bechtold et al. used an interplay between mortality and density in their gross growth model to capture an effect that Nelson (1963) captured with a quadratic density term. Namely, that density increases to a threshold at which competition-related mortality begins to increase causing growth to decline. Tree based models automatically attempt to accommodate quadratic terms (and higher order terms and interactions if necessary) so we did not include mortality in our gross growth models. Instead, we accepted a quadratic density term in the model if it helped explain gross growth.

Our technique for analyzing Data #1 and Data #2 can be subsumed under 2 broad headings. 1) Since our objective was to determine whether there is a change in growth rates from cycle 4 to cycle 5 (i.e., a cycle effect), we explained as much structure as possible by the covariates before testing for a cycle effect, so we accepted a slight overfit of the data. 2) We performed a test of the cycle effect on growth after as much data structure as possible had been explained by the other covariates. The idea of removing confounding effects of covariates from the estimation of an effect of one particular variable using tree based procedures was also explored by Siu et al. (1985), though they used a tree growing procedure and testing technique different from ours.

To fit a tree model using all covariates except cycle, one tests the predictive capability of the model to determine how much the overfit might be and then scales back the model if it appears too much. This judgment is somewhat subjective, but the criteria used to make it are described below.

1. Tree based regression is carried out using the appropriate model.
2. The tree model is pruned using several criteria.
   a. Cross validation is performed on a tree sequence.
      i. The tree sequence is obtained by defining a vector of cost-complexity parameters (Clark and Pregibon 1993) δ, and pruning the tree model to a size determined by a value in δ so that $D_\delta(T)$ in equation (4) is a minimum. At each tree size of the sequence, a m-fold cross validation is performed, where m is determined to make subset size around an average of 25 to 40. Subsets are formed by randomly subdividing the observations into m groups of roughly equal size. A plot of a typical cross validation analysis is shown in figure 4.
      ii. The above step is carried out 500 times and at each iteration the size of tree where the mini-

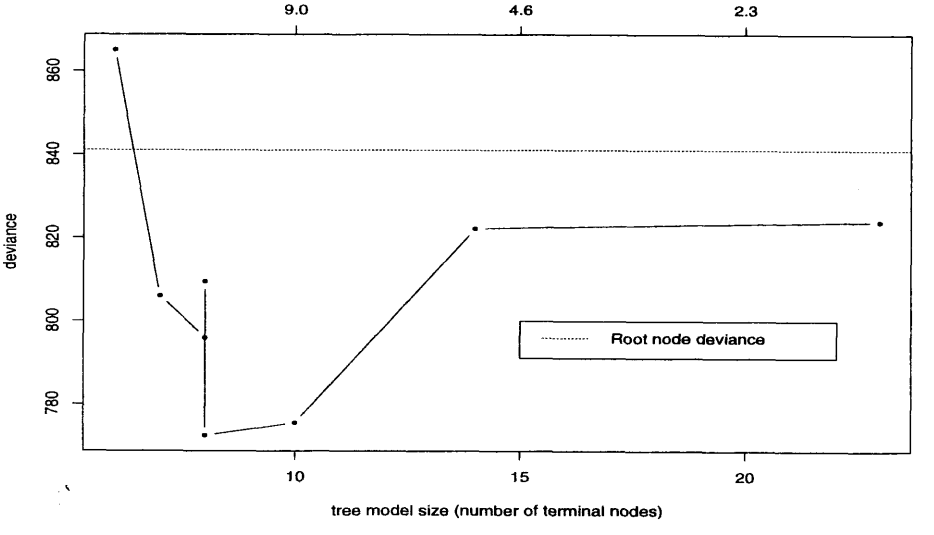


Figure 4. Shortleaf pine only, net growth model (Data #1). A 6-fold cross validation with δ =1, 5, 9, ... 29, and the subsets chosen so that the first comprises the first 29 observations, the 2nd the next 29 observations, and so on with the last two subsets comprising 28 observations each. The root node deviance is the deviance of a one node tree (i.e., the sum of squares about the mean of all observations).

mum deviance occurred is recorded with its corresponding minimum deviance. The distribution of tree model sizes where minimum deviance is attained can be viewed via a plot of tree model sizes versus corresponding minimum deviance. The mode of the distribution of these tree sizes is easily obtained. A typical plot is shown in figure 5, which suggests that an 8 node tree is appropriate (although 13 nodes could be considered if overfitting were not a major concern). One could interpret figure 5 as suggesting a 3 node tree, but it is unlikely that such a small tree model will capture the structure inherent in the data.

iii. Several plots of the type shown in figure 4 are viewed to determine the deviance fluctuation for changing tree model sizes. It is not unusual for the deviance curve to be fairly flat for a range of tree sizes, although in some cases the deviance fluctuates wildly with changing tree model size making interpretation difficult. Figure 4 suggests an 8 to 10 node tree.

b. A deviance curve for increasing tree model sizes is plotted (figure 6). This plot shows that deviance decreases sharply as tree size increases from 1 to around 10 terminal nodes and then begins to level out, which indicates that adding more nodes is having less effect on the overall tree deviance.

c. Terminal nodes containing the most observations are identified and a linear regression of Y on the covariates is run for each node. Two models are used and in each variables are transformed (Bechtold et al. and Ruark et al.) to obtain approximate linearity among response and covariates. Note that mortality is not used in any of our

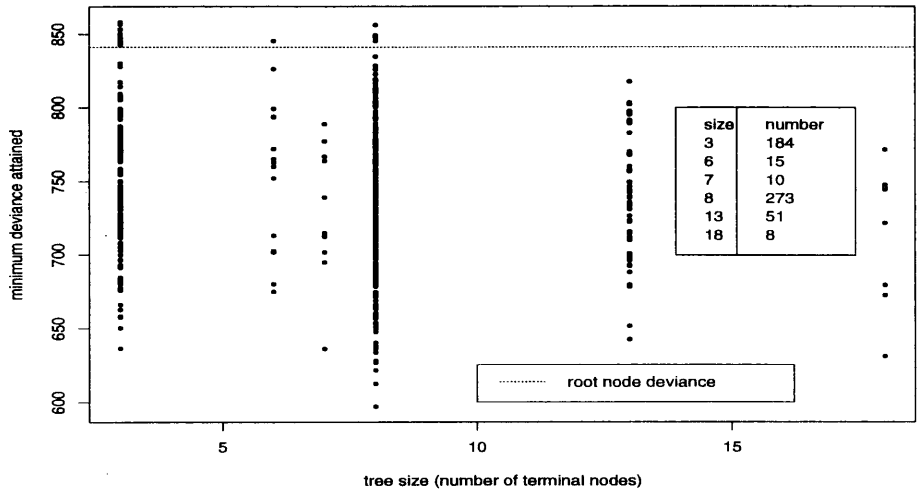


Figure 5. Shortleaf pine only, net growth model (Data #1). Minimum deviance versus corresponding tree model size for 500 iterations of 6-fold cross validations. The legend provides the number of times minimum deviance occurs for a particular size of tree model.

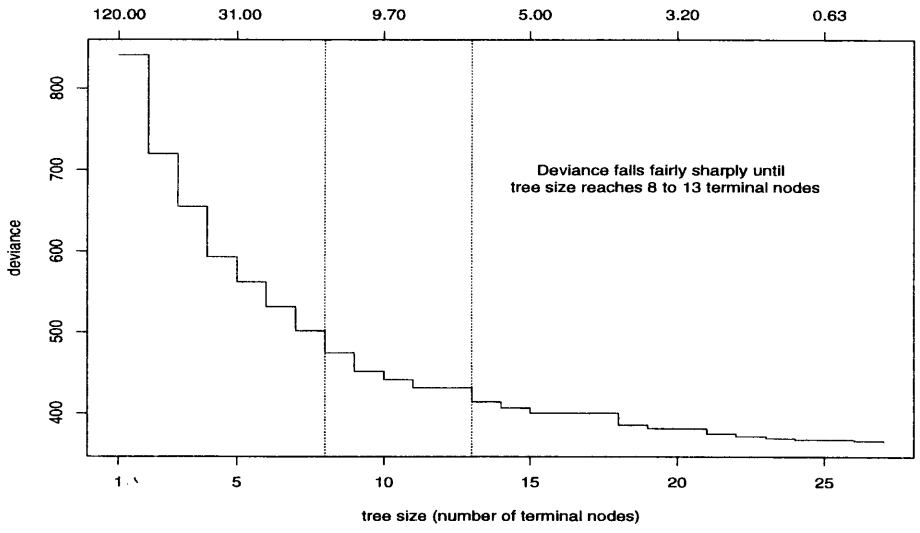


Figure 6: Shortleaf pine only, net growth model (Data #1). Deviance curve for increasing tree sizes. The 2 vertical lines indicate an area of the graph where the curve tends to flatten out, though this is a visual interpretation and somewhat arbitrary.

models, and, for our net growth model, the response is log-transformed for both data sets but the age covariate in Data #1 is untransformed. The other covariates are transformed in the same way as the gross growth model. First, a regression is run on one predictor at a time. If any of the predictors are significant, it may indicate that the tree should be allowed to grow larger to explain this additional structure at the tested node. Another regression is run within each node with all predictor variables in the model. If this model is significant, it may be that a limitation in the tree based model approach has been encountered since the splits are restricted to one variable at a time; splits on significant linear combinations of covariates are not allowed in our tree growing algorithm. In our analyses, if regression results show significance at terminal nodes, depending on how many associations are found between response and covariates, we force the model to split at these nodes, which allows the tree model to grow larger to capture this structure. Analyses in the results section are marked with a " * " if this occurred.

3. A satisfactorily pruned tree is constructed by considering the information obtained from the above steps. A plot of such a tree is in figure 7. The selected tree was pruned to 13 terminal nodes. For this particular tree model size, there are no covariate linear combinations that significantly explain the observations in either of the 2 most populated terminal nodes.

4. We use a nonparametric procedure that does not require the response to be normally distributed to test for cycle effect (i.e., tree growth in cycles 4 and 5 being significantly different). Testing for a cycle effect differs from earlier analyses. As noted, we do not include the cycle indicator variable in our tree based models but construct a tree model to explain the overall relationship between response and covariates without the cycle term. This model places observations into homogeneous groups (i.e., clusters of observations having similar values for the covariates and, within these clusters, there is no strong evidence of a relation between response and covariates). Next, we test for cycle effect within the clusters employing a nonparametric permutation procedure (Mielke and Iyer 1982). This test is analogous to a block design analysis of variance where terminal tree nodes are blocks and cycle is the treatment factor. The permutation test procedure calculates a p-value for the null hypotheses H_0: cycle effect=0.

5. Finally, to compare this with more traditional techniques, all subsets regression by leaps and bounds

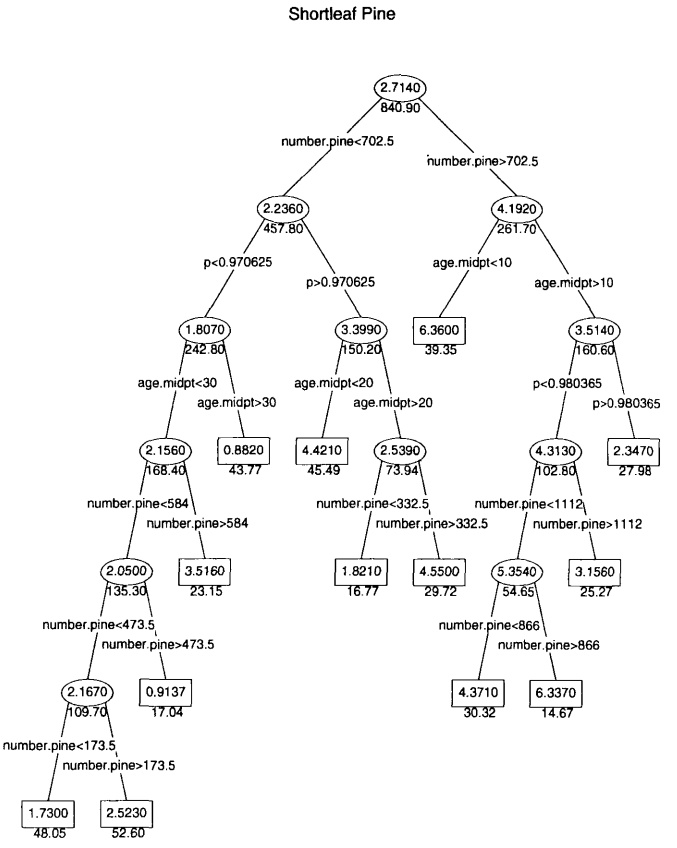


Figure 7: Shortleaf pine only, net growth model (Data #1). A 13 node tree. Rectangles are terminal nodes. The number inside the rectangle/oval is the predicted value (mean value) at that node. The number below the rectangles/ovals is the deviance at that node.

(Furnival and Wilson 1974) was run on the pool of covariates created by the transformed covariates listed above (again excluding the term for cycle), their squares, and all 2-way interactions. The best model as ranked by Mallow's *Cp* was obtained and the deviance was computed.

Thirty-three tree based models, corresponding to the 33 analyses reported in this paper, were developed using the above steps. The results of these analyses are presented in the next section. We discuss results that are particularly interesting and present others in a table.

Results and Discussion

Results are in tables 2 through 5. Tree model sizes and their deviance are reported along with the results of the

permutation test of cycle effect=0. The adequacy of the tree model is judged as G=good, F=fair, and P=poor. Poor models were grown larger at populated terminal nodes to account for additional data structure at these nodes (see the final note of item 2 above). This judgment is subjective and is based on the strength of associations between response and covariates at highly populated terminal nodes. A final column shows the deviance and degrees of freedom of a linear model selected according to item 5 in the Methods section. This deviance is in arithmetic units and is corrected for log bias. Even though deviance cannot be accurately compared between tree based models and linear models, our tree based models have lower deviance than corresponding linear models. But linear models seem adept at capturing the structure inherent in the analyzed data sets.

The 6 analyses for gross growth of DATA #1 offer conclusions regarding cycle effect that generally agree with Bechtold et al. although our evidence for a decline in gross growth of slash pine are more marginal. The net

Table 2. Data #1: Gross growth analysis.

Tree/plot type	Tree size	Tree model deviance (d.f.)	Tree model H_0: cyc.effect =0 p-value	Adequacy of tree model*	Linear model deviance (d.f.)
Loblolly-pine only**	23	924 (316)	0.0006	G/F	1190 (331)
Loblolly-all trees	9	1429 (330)	0.0001	F	1555 (332)
Shortleaf-pine only	14	268 (158)	0.0302	G	390 (165)
Shortleaf-all trees	15	389 (157)	0.0080	F	556 (165)
Slash-pine only	10	591 (157)	0.0579	G	757 (160)
Slash-all trees	7	834 (160)	0.0469	G	922 (161)

**G=Good, F=Fair.*

***Further growth was required at one or more terminal nodes.*

Table 3. Data #1: Net growth analysis.

Tree/plot type	Tree size	Tree model deviance (d.f.)	Tree model H_0: cyc.effect =0 p-value	Adequacy of tree model*	Linear model deviance (d.f.)
Loblolly-pine only	** 13	1486 (326)	0.0732	G	1694 (328)
Loblolly-all trees	9	2013 (330)	0.0006	G	2146 (331)
Shortleaf-pine only	13	414 (159)	0.0207	G	573 (166)
Shortleaf-all trees	6	747 (166)	0.0362	G	798 (166)
Slash-pine only	10	686 (157)	0.2183	G	832 (160)
Slash-all trees	6	938 (161)	0.1471	G	1006 (161)

**G=Good.*

***Further growth was required at one or more terminal nodes.*

growth analysis of Data #1 revealed some interesting conclusions; the net growth of slash pine did not change significantly from cycle 4 to cycle 5. Since Bechtold et al. did not analyze net growth, these findings are not contradictory.

Our analyses of gross growth of Data #2 showed some unexpected departures from the conclusions of Ruark et al. Georgia longleaf *PSG5* gross growth did not significantly change (p-value=0.1063) from cycle 4 to cycle 5. Alabama loblolly *PSG1* gross growth did not appear to change either (p-value=0.2876). Ruark et al. showed Alabama loblolly *PSG5* gross growth as not significantly changing with cycle whereas our analysis showed significance (p-value=0.0360). Also in our study, Alabama shortleaf *PSG1* gross growth did not show a significant change with cycle (p-value=0.0869), but the p-value was sufficiently low so there is no serious departure from Ruark et al.

Our net growth analysis of Data #2 revealed a significant change of growth with cycle in all tree types with the exception of Georgia longleaf *PSG1*, which showed growth rates between cycles being so similar that the p-value is very close to 1. The program that we used rounds the p-value up to 1.00 after it exceeds a certain

Table 4. Data #2: Gross growth analysis.

Tree/plot type State	Tree size	Tree model deviance (d.f.)	Tree model H_0: cyc.effect =0 p-value	Adequacy of tree model*	Linear model deviance (d.f.)
Georgia					
Loblolly - PSG1	16	2693 (784)	0	F	3137 (791)
Loblolly - PSG5	24	713 (663)	0	G	916 (679)
Shortleaf - PSG1	28	478 (332)	0	G	784 (353)
Shortleaf - PSG5	11	169 (296)	0	G/F	219 (301)
Longleaf - PSG1	18	156 (223)	0.0324	G/F	241 (233)
Longleaf - PSG5	12	75 (207)	0.1063	G/F	106 (212)
Slash - PSG1	37	849 (461)	0.0007	G	1367 (488)
Slash - PSG5	31	275 (404)	0.0449	G	427 (424)
Alabama					
Loblolly - PSG1**	26	1363 (456)	0.2876	G/F	1982 (473)
Loblolly - PSG5	13	358 (451)	0.0360	G	432 (456)
Shortleaf - PSG1	16	623 (239)	0.0869	G/F	853 (249)
Shortleaf - PSG5**	17	92 (220)	0.0019	G	129 (229)
Longleaf - PSG1	7	126 (135)	0.0069	G	149 (134)
Longleaf - PSG5	6	27 (132)	0.0059	G	33 (131)

**G=Good, F=Fair.*

***Further growth was required at one or more terminal nodes.*

threshold based on the test statistic and a skewness coefficient. We are comfortable stating that the p-value is >0.9.

Tables 6 through 8 detail tree models that produced results somewhat different from prior analyses. Details are shown for Georgia longleaf *PSG*5 gross growth, Alabama loblolly *PSG*1 gross growth, and the net growth model of Georgia longleaf *PSG*1.

Sensitivity Assessment

We assessed how sensitive our results were to the specific rule used to find a tree model (ignoring cycle), and whether or not the distribution of the covariates changes from one cycle to the next. When the same data set is used to determine a model and conduct inferences on coefficients, certain biases are expected that are similar to those in typical multiple comparisons problem when testing several hypotheses. The reported p-values are sensitive to the model selection procedure and are unlikely to be exact. With regard to subset selection in regression, A. J. Miller (1990) states, "When the selected model is the best-fitting in some sense, conventional fitting methods give estimates of regression coefficients which are usually biased in the direction of being too large." That is, even when a predictor variable is randomly generated without relation to a response, it will be significant more frequently than expected.

The tree based methodology used in our analysis of the Bechtold et al. and Ruark et al. data sets is an approximate inference procedure in the sense that the p-values reported by the analyses are not expected to be exact. However, for the procedure to be reliable for our application, tests of the effect of cycle on growth rates conducted using a nominal α level should have an actual α level that is close to the nominal value. To assess this closeness, we conducted a small simulation study that is summarized below.

We used the Bechtold et. al loblolly pine data set as a model for generating simulated covariate data (not including the cycle variable). The response variable, growth rate, was generated by using a linear model fitted to the actual data set. This linear model was obtained using all subsets regression by leaps and bounds (Furnival and Wilson 1974) on the pool of covariates (transformed as per Bechtold et. al.), their quadratic terms, and 2-way interactions. The errors used were independently and identically distributed random normal variables with mean zero and

Table 5. Data #2: Net growth analysis.

Tree/plot type State	Tree size	Tree model deviance (d.f.)	Tree model H_0: cyc.effect =0 p-value	Adequacy of tree model*	Linear model deviance (d.f.)
Georgia					
Loblolly - PSG1	34	3877	0	G	4931
		(766)			(789)
Shortleaf - PSG1	12	1201	0	G	1484
		(348)			(352)
Longleaf - PSG1**	12	272	> 0.9	G	298
		(229)			(233)
Slash - PSG1	22	1404	0.0002	G	1736
		(476)			(487)
Alabama					
Loblolly - PSG1**	17	2412	0.0106	F	2895
		(465)			(472)
Shortleaf - PSG1	9	1417	0.0003	G	1830
		(246)			(246)
Longleaf - PSG1	11	144	0.0033	G	187
		(131)			(135)

**G=Good, F=Fair.*

***Further growth was required at one or more terminal nodes.*

Table 6. Tree model details for Georgia longleaf PSG5 gross growth. Eight of the 12 nodes have y4.m>y5.m, but the remaining 4 nodes, including the highly populated node 1, have y4.m< y5.m. Thus, the evidence for a growth decline is not very strong and the p-value for H_0:cycle effect=0 is 0.1063.

node	n4	n5	y4.m	y5.m	y.m	y4.sd	y5.sd	y.sd
1	31	14	0.455	0.509	0.472	0.229	0.192	0.217
2	36	17	1.150	0.985	1.098	0.375	0.376	0.380
3	25	23	1.642	1.399	1.526	0.610	0.437	0.542
4	5	6	2.198	2.281	2.243	0.727	1.127	0.921
5	4	4	3.334	3.280	3.307	0.774	0.912	0.784
6	9	2	2.643	2.057	2.537	1.139	0.230	1.048
7	3	2	3.511	3.463	3.492	0.788	0.032	0.558
8	6	2	1.995	1.562	1.887	1.202	1.474	1.176
9	2	4	3.464	2.504	2.824	0.643	0.361	0.637
10	7	5	3.859	3.641	3.768	0.712	1.494	1.049
11	3	4	1.959	2.070	2.023	0.380	0.665	0.522
12	2	3	1.005	1.071	1.044	0.443	0.508	0.423

node=terminal node number.

n4, n5=number of cycle 4 and cycle 5 observations in node.

y4.m, y5.m=mean of cycle 4 and cycle 5 observations in node.

y.m=mean of all observations in node (also the predicted value for that node if cycle is ignored).

y4.sd, y5.sd, y.sd=corresponding standard deviations.

Table 7. Tree model details for Alabama loblolly PSG1 gross growth. Of the 26 nodes, 3 do not have any observations from cycle 4. Sixteen of the remaining 23 nodes have y4.m>y5.m indicating a growth decline while the remaining 7 nodes have y4.m <y5.m. Nodes 14, 21, 22, and 23 have some high standard deviations indicating possible influence of outliers. We reran the permutation test after eliminating these nodes and obtained a p-value=0.2444, indicating that these nodes had only a minimal influence on the permutation test results.

node	n4	n5	y4.m	y5.m	y.m	y4.sd	y5.sd	y.sd
1	15	8	0.557	0.746	0.623	0.326	0.395	0.354
2	32	26	1.419	1.401	1.411	0.831	0.756	0.791
3	3	8	1.293	1.131	1.175	0.727	0.430	0.490
4	44	30	2.244	1.975	2.135	0.927	0.849	0.900
5	17	32	2.782	2.288	2.460	1.426	1.042	1.198
6	4	2	4.471	4.122	4.355	1.428	0.116	1.122
7	7	19	3.849	2.152	2.609	1.156	1.175	1.380
8	4	7	5.253	4.517	4.785	1.843	1.810	1.767
9	5	5	2.498	2.934	2.716	1.271	1.448	1.305
10	10	11	3.047	3.690	3.384	2.006	2.586	2.294
11	3	10	6.264	5.765	5.880	1.081	2.579	2.287
12	0	9	NA	3.780	3.780	NA	1.615	1.615
13	0	6	NA	2.148	2.148	NA	0.947	0.947
14	3	2	9.341	5.490	7.801	6.291	2.579	5.089
15	1	11	2.204	3.719	3.593	NA	0.883	0.949
16	3	3	5.163	3.369	4.266	0.534	1.966	1.620
17	2	3	6.652	8.567	7.801	3.847	1.936	2.583
18	10	44	2.678	2.275	2.349	1.027	1.153	1.132
19	5	8	3.039	2.417	2.656	1.074	0.744	0.898
20	6	14	2.937	4.191	3.815	0.622	1.261	1.240
21	4	1	12.680	7.838	11.712	8.353	NA	7.551
22	2	5	5.769	9.703	8.579	1.324	5.044	4.575
23	6	8	8.233	6.041	6.980	5.461	2.138	3.899
24	1	5	3.640	3.317	3.371	NA	1.048	0.946
25	3	10	5.727	5.638	5.659	2.553	2.475	2.384
26	0	5	NA	3.053	3.053	NA	2.017	2.017

variance equal to the mean squared error from the leaps and bounds linear model. The generated response was the logarithm of the growth rate, so it was untransformed to obtain the true gross growth values. The cycle variable was generated by correlating it to a function of covariates (using the Bechtold et. al. loblolly pine data set as a model), but independent of the response. In different simulation runs, the degree of correlation between cycle and the covariates was varied so that 3 different scenarios were represented; no correlation, small correlation, and moderate correlation. The correlation between response and cycle would be present only through the covariates. Under these conditions, gross growth should be independent of cycle once the covariates were accounted for. In other words, the null hypothesis of no cycle effect is true for the simulated data.

At each iteration (1000 total) of the simulation, data was generated, a tree model was constructed (based on the techniques described in the Methods section), and a test for cycle effect in the terminal nodes was performed with the permutation test. Some subjective judgement was used when analyzing the actual data sets to select a tree model. Subjective judgement was impossible in the automated computer simulations, so with each iteration we grew a tree model using the S-Plus default tree growing algorithm (Clark and Pregibon 1993) and then pruned to a deviance approximately equal to a 20% increase in the deviance of the fully grown default tree model. In nodes with more than 30 observations, existence of linear relations between response and covariates was checked as described in the Methods section. If significant relationships (p-value≤.08) were found, these nodes were further split in proportion to the number of significant covariates. We felt that this procedure would sufficiently mimic what was done during the actual data analysis.

We found that when the covariates were unrelated to cycle, the reported p-values agreed with the actual p-values. When there was a moderate correlation between cycle and the covariates, the reported p-values were smaller by a factor of 2 or 3. We never found the actual p-value exceeding 0.05 when the reported p-value was 0.01 or smaller.

Although the simulation study was not exhaustive, it suggests that caution should be used when interpreting p-values based on the tree regression methodology. We would not claim a significant cycle effect unless the associated p-value was 0.01 or smaller.

Summary

The Bechtold et al. and Ruark et al. data sets were analyzed using tree based regression models and nonparametric permutation tests for cycle effect. Of importance is the screening criteria used to select the sample plots of pine stands. Actual populations of inference to which these results apply are unknown because sample plots were carefully screened to make the samples for the 2 periods comparable. Our results generally agreed with those of the previous authors.

Tree based models make few assumptions and provide a viable alternative to standard regression methodology. The methodology used here can also be used when the significance of a specific covariate is in question. This methodology is a useful, robust (i.e., not restricted to model assumptions that include assumed model forms and assumed distributional forms of the data), and interpretable approach for examining whether change has occurred over time in variables of interest in large-scale survey data sets.

Table 8. Tree model details for Georgia longleaf PSG1 net growth. Node 9 has no stands from cycle 4. Of the remaining 11 nodes, 6 have y4.m>y5.m while 5 have y4.m<y5.m, which suggests that there is no evidence of a growth decline. The p-value is>0.9.

node	n4	n5	y4.m	y5.m	y.m	y4.sd	y5.sd	y.sd
1	15	5	1.118	1.011	1.092	0.430	0.251	0.389
2	13	7	1.834	1.901	1.858	0.732	1.178	0.882
3	21	11	0.465	0.497	0.476	0.258	0.193	0.235
4	14	13	0.942	0.915	0.929	0.686	0.573	0.622
5	9	6	1.733	1.302	1.561	0.989	1.569	1.219
6	48	29	2.757	2.367	2.610	1.257	1.336	1.293
7	3	6	0.470	1.769	1.336	0.440	0.766	0.915
8	3	3	3.662	4.075	3.868	2.281	1.308	1.678
9	5	0	5.846	NA	5.846	1.939	NA	1.939
10	3	3	2.939	2.240	2.590	1.521	0.963	1.201
11	6	3	3.915	5.422	4.418	1.906	1.645	1.875
12	8	7	2.837	2.810	2.825	1.547	1.096	1.308

Acknowledgment

The authors wish to thank Dr. Jennifer Hoeting for her helpful comments on an earlier version of this manuscript.

Literature Cited

Bechtold, W. A., Ruark, G. A. and F. T. Lloyd (1991), **Changing Stand Structure and Regional Growth Reductions in Georgia's Natural Pine Stands.** Forest Science 37, 703-717.

Breiman, L., Friedman, J. H., Olshen, R. A. and C. J. Stone (1984), **Classification and Regression Trees.** Monterey, California: Wadsworth, Inc.

Chaudhuri, P., Lo, W., Loh, W. and C. Yang (1995), **Generalized Regression Trees.** Statistica Sinica 5, 641-666.

Ciampi, A., Lawless, J. F., McKinney, S. M. and K. Singhal (1988), **Regression and Recursive Partition Strategies in the Analysis of Medical Survival Data.** J. Clinical Epidemiol. 41, No. 8, 737-748.

Clark, L. A. and D. Pregibon (1993), **Tree based models. In Statistical Models** in S, J. M. Chambers and T. J. Hastie, ed., New York: Chapman and Hall.

Furnival, G. M. and Wilson, R. W. Jr. (1974), **Regressions by Leaps and Bounds.** Technometrics 16, 499-511.

Le May, V. M. (1994), **Estimating the probability and amount of decayed wood in standing trees, IUFRO Conference: Simplicity and Efficiency in Sampling and Non-Commodity Use of Surveys.** Ascona, Switzerland. May 1994.

Mardia, K. V., Kent, J. T. and J. M. Bibby (1979), **Multivariate Analysis.** San Diego: Academic Press.

Mielke, P. W. and H. K. Iyer (1982), **Permutation Techniques for Analyzing Multi-Response Data from Randomized Block Experiments.** Commun. Statist.-Theor. Meth. 11, 1427-1437.

Miller A. J. (1990), **Subset Selection in Regression.** New York: Chapman and Hall.

Morgan, J. N. and J. A. Sonquist (1963), **Problems in the Analysis of Survey Data, and a Proposal.** Journal of the American Statistical Association 58, 415-434.

Nelson, T. C. (1963), **Basal Area Growth of Natural Loblolly Pine Stands.** USDA For. Serv. Southern Forest Experiment Station Res. Note SE-11, 4 pp.

Ouyang, Z., Schreuder, H. T. and H. G. Li (1992), **A reevaluation of the Growth Decline in Georgia and Georgia-Alabama.** Proc. 1991 Kansas State Univ. Conf. on Appl. Statistics in Agriculture April 28-30, 1991, Manhatten, Kansas, 54-61.

Ruark G. A., Thomas, C. E., Bechtold, W. A. and D. M. May (1991), **Growth Reductions in Naturally Regenerated Southern Pine Stands in Alabama and Georgia.** South. J. Appl. For. 15, 73-79.

Schreuder, H. T. and C. E. Thomas (1991), **Establishing Cause-Effect Relationships Using Forest Survey Data.** Forest Science 37, 1497-1525 (includes discussion).

Sheffield, R. M. and N. D. Cost (1987), **Behind the Decline.** J. For. 85, 29-33.

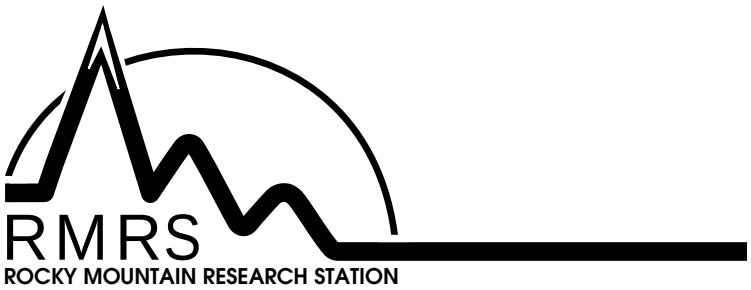
Sheffield, R. M., Cost, N. D., Bechtold, W. A. and J. P. McClure (1985), **Pine Growth Reductions in the Southeast.** USDA For. Serv. Res. Bull. SE-83, 112 pp.

Siu, C. O. and A. F. Andrews (1985), **Piecewise Linear Tree-structured Regression with an Application for Covariance Analysis.** In ASA Proc. Statist. Comput. Sect. 215--219. Alexandria, VA: Amer. Statist. Assoc.

Temkin, N. R., Holubkov, R., Machamer, J., Winn, H. R. and S. S. Dikmen (1995), **Classification and Regression Trees (CART) for Prediction of Function at 1 Year Following Head Trauma.** J. Neurosurg 82, 764-771.

Verbyla, D. L. (1987), **Classification Trees: A New Discrimination Tool.** Can. J. For. Res. 17, 1150-1152.

Zahner, R., Saucier, R. J. and R. K. Meyers (1989), **Tree-ring Model Interprets Growth Declines in the Southeastern United States.** Can. J. For. Res. 19, 612-621.



The Rocky Mountain Research Station develops scientific information and technology to improve management, protection, and use of the forests and rangelands. Research is designed to meet the needs of National Forest managers, Federal and State agencies, public and private organizations, academic institutions, industry, and individuals.

Studies accelerate solutions to problems involving ecosystems, range, forests, water, recreation, fire, resource inventory, land reclamation, community sustainability, forest engineering technology, multiple use economics, wildlife and fish habitat, and forest insects and diseases. Studies are conducted cooperatively, and applications may be found worldwide.

Research Locations

Flagstaff, Arizona
Fort Collins, Colorado*
Boise, Idaho
Moscow, Idaho
Bozeman, Montana
Missoula, Montana
Lincoln, Nebraska
Reno, Nevada
Albuquerque, New Mexico
Rapid City, South Dakota
Logan, Utah
Ogden, Utah
Provo, Utah
Laramie, Wyoming

*Station Headquarters, 240 West Prospect Road, Fort Collins, CO 80526

Federal Recycling Program  Printed on Recycled Paper