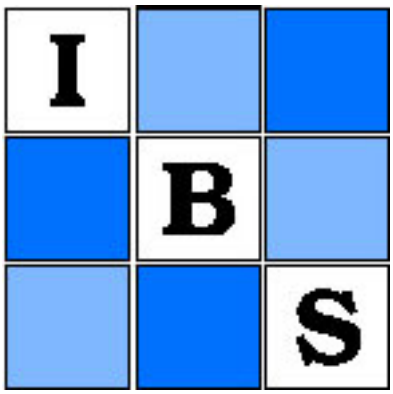

An Adaptive Analysis of Covariance Using Tree-Structured Regression

G. L. GADBURY, H. K. IYER, and H. T. SCHREUDER

In this article, we propose an adaptive procedure for testing for the effect of a factor of interest in the presence of one or more confounding variables in observational studies. It is especially relevant for applications where the factor of interest has a suspected causal relationship with a response. This procedure is not tied to linear modeling or normal distribution theory, and it offers a valuable alternative to traditional methods. It is suitable for applications where a factor of interest is categorical and the response is continuous. Confounding variables may be continuous or categorical. The method is comprised of two basic steps that are performed in sequence. First, confounding variables alone (i.e., without the factor of interest) are used to group observations into subsets. These subsets have the property that, when restricted to a subset, there is little or no remaining variation in the response that is attributable to the confounding variables. We then test for the factor of interest within each subset of observations. We propose to implement the first step using a technique that is generally referred to as tree-structured regression. We use a nonparametric permutation procedure to carry out the second step. The proposed method is illustrated through an analysis of a U. S. Department of Agriculture (USDA) Forest Service data set and an air pollution data set.

Key Words: Classification and regression trees; Nonparametric; Permutation test; Tree.

1. INTRODUCTION

A common problem faced by scientists in many fields is the assessment of the effect that a factor has on a response using data collected in observational studies. The problem can become complicated if it is suspected that other variables (confounding variables) are also affecting the response in addition to, or instead of, the factor in question. In this case, one needs to study the dependence of the response on the factor of interest conditional on the confounding variables. If data are available on the confounding variables, then a customary approach to this problem is the use of analysis of covariance techniques.

Gary Gadbury is an Assistant Professor in the Department of Mathematics and Statistics, University of Missouri at Rolla, Rolla, MO (E-mail: gadburyg@umr.edu). Hari Iyer is a Professor in the Department of Statistics, Colorado State University, Fort Collins, CO 80523. Hans Schreuder is a Mathematical Statistician at the USDA Forest Service, Rocky Mountain Research Station, Fort Collins, CO 80526.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 7, Number 1, Pages 42–57

Traditional analysis of covariance assumes that the conditional expectation of the response, given all the covariates, is of a known form. Typically, a linear model relationship is assumed. Under normality assumptions, exact tests are available for the effect of the test factor conditional on the covariates.

In practice, the form of the relationship between the response and covariates is unknown and could be complex. This was true of some U.S. Department of Agriculture (USDA) Forest Service Forest Inventory and Analysis (FIA) data sets that motivated the method described in this article. These data were obtained from surveys of forest resources over time. Growth rates for pine stands in the southeastern United States were recorded in consecutive 10-year periods in addition to several covariates that represented structural characteristics of the pine stands. The interest was in determining if tree growth rates changed from one 10-year period to the next after accounting for changes in stand structure.

Since the true form of the relationship between the response and covariates was unknown, the investigation required two steps: (1) a model-building step and (2) a testing step. In general, P-values calculated using such adaptive procedures will not be exact. Bechtold, Ruark, and Lloyd (1991) analyzed these data using traditional analysis of covariance. They transformed covariates as suggested by diagnostics in order to obtain a good model and then tested for a change in tree growth rates (conditional on the covariates) using normal distribution theory.

In this article, we propose an alternative adaptive procedure to investigate the effect of a factor on a response, conditional on covariates. This procedure requires minimal assumptions regarding the distributional form of the data and minimal assumptions regarding the form of the model relating covariates to the response. Our method is suitable as an analysis of covariance procedure for applications when the response variable is continuous and when the factor in question is categorical. Confounding variables can be categorical or continuous. The method is particularly suited to applications where there is a suspected cause–effect relationship between the factor of interest and the response.

We use tree-structured regression for the model-building step and a permutation test for the testing step. After describing our proposed method, we illustrate the method using an USDA FIA data set. Results of a simulation study reported elsewhere (Gadbury, Iyer, Schreuder, and Ueng 1998) indicate that the type I error rate of the proposed procedure was sufficiently close to the nominal values to make the procedure viable for practical use. We then demonstrate the flexibility of the procedure by using it with the air pollution data that were analyzed in McDonald and Schwing (1973). We conclude with a discussion.

2. PROPOSED METHOD

Consider a model of the form

$$Y = f(X_1, \ldots, X_{p-1}, X_p) + \epsilon = f(\boldsymbol{X}, X_p) + \epsilon, \tag{2.1}$$

where Y indicates a continuous response variable, $\boldsymbol{X} = (X_1, \ldots, X_{p-1})$ are confounding variables that we will simply refer to as covariates, and ϵ is a random error component. We

refer to the factor of interest, or the test factor, as X_p. First, we use covariates to construct a predictive model. This model divides data into, say, K subsets of observations such that, within each subset, there is very little remaining variation in the response that can be explained by $\boldsymbol{X}$. Variation in Y predominantly occurs across subsets. We then determine if the test factor, X_p, significantly reduces any remaining variation in Y within each subset.

The two steps we use—building a model relating the response to the covariates and then including the test factor to determine its effect on the response given the covariates—are sequential steps. In standard analysis of covariance, the test factor is included in the initial model building step. If the model-building step involves a variable selection procedure and X_p is included in the pool of predictors, there is a possibility that the test factor will mask the role of one or more observed covariates in explaining the variation in response. Since our null hypothesis is that the test factor has no effect on the response variable, we attempt to find a best model without X_p so that the pattern in the response variable is explained by patterns in the covariates, even if a slight overfit of the data occurs. With this model, we can then feel reasonably certain that any possible effects of observed confounding variables, $\boldsymbol{X}$, have been considered before testing for the effect of X_p. This is not to say that there might not be other unobserved covariates that might also explain the pattern in the response. The potential for unobserved covariates and resulting hidden biases is a characteristic of observational data that will not be considered further here. See Rosenbaum (1995) for details on unobserved covariates and resulting hidden biases.

2.1 DEVELOPING A MODEL TO PREDICT Y USING $\boldsymbol{X}$

We use tree-based models or tree regression to fit the response variable to covariates $\boldsymbol{X}$. Tree-based models can be used to study both classification and regression problems and, thus, are often referred to as classification and regression trees (CART) (Breiman, Friedman, Olshen, and Stone 1984).

We use tree regression to group observations into subsets such that, within each subset, there is very little remaining relationship between the response variable and confounding variables. (See Breiman et al. (1984) for details of classification and regression trees and Clark and Pregibon (1993) for a description of the S-Plus implementation of CART that we used.) We describe the steps that we used to construct the tree model and include an overview of certain tree regression procedures, as needed, along the way.

The Model-Building Step

Our model-building procedure is described below and involves two basic steps, (1) constructing a tree regression model and (2) refining the model.

(1) Tree-structured regression is carried out using all variables except X_p. Initially, all of the n observations are in one group, say G_0, called the root node of the tree model. Let the mean of the n response values in G_0 be $\bar{y}_0$. Then $\bar{y}_0$ is a prediction for observations in G_0.

The sums of squared deviations about the mean of observations in a node is called the

deviance for a node. We use the term deviance to remain consistent with terms in S-Plus (Clark and Pregibon 1993). The deviance for a node G_k is given by

$$D(G_k) = \sum_{i=1}^{n_k} (y_i - \bar{y}_k)^2,$$

where n_k is the number of observations in G_k and $\bar{y}_k$ is the mean of the response values in G_k. The observations in G_0 are divided into two groups or nodes, G_1 and G_2, on the basis of a selected value of a covariate, say x_0. Observations with a value of this covariate less than or equal to x_0 are grouped into G_1, and observations with a value greater than x_0 are grouped into G_2. The tree regression procedure selects this covariate and this value, x_0, so that the quantity

$$D(G_0) - D(G_1) - D(G_2)$$

is maximized, i.e., a maximum reduction in deviance is obtained. The algorithm continues in this manner to divide data in one node into two nodes until one of two termination criteria is met. When a termination criterion is satisfied for a particular node, then that node will not divide further and is called a terminal node of the model. The size of the model refers to the number of terminal nodes in the model, and the deviance of the model is the sum of all terminal node deviances. The two termination criteria for a node, say G_k, are

(a) The number of observations in G_k, n_k, is too small to justify further divisions, i.e., $n_k \leq n_{\min}$.

(b) The response values in G_k are nearly homogeneous when compared with observations in the root node G_0. This can be controlled by specifying γ in the condition $D(G_k) \leq \gamma D(G_0)$. This criterion is somewhat analogous to determining how many higher order terms and interactions to include in a linear regression model.

The two values $n_{\min}$ and γ are tuning constants that determine how many terminal nodes the tree regression procedure will produce. The second criterion involving γ does not incorporate the number of observations in G_k. If some test of deviance was of interest, one might opt to change the criteria to compare mean square error of G_k with that of the root node. However, no testing is done to determine if a division of data was warranted at a particular node. The termination criteria are set to favor overfitting, rather than underfitting, the data, i.e., $n_{\min}$ is set to a small number (10 is the default) and γ is set to a small number (0.01 by default). One could set $n_{\min} = 2$ and $\gamma = 10^{-10}$, thus producing a model that fits the data exactly, but this will generally entail unnecessary computation. Usually with the default values of $n_{\min}$ and γ, many divisions of nodes may not have been necessary and the resulting tree model size will likely be larger than needed. So the model is refined using a technique called pruning the tree model. This technique is analogous to backward stepwise selection in linear regression.

(2) A tree model can be pruned by observing which divisions of the data do not contribute to a large reduction in deviance. The pruning function is automated as follows.

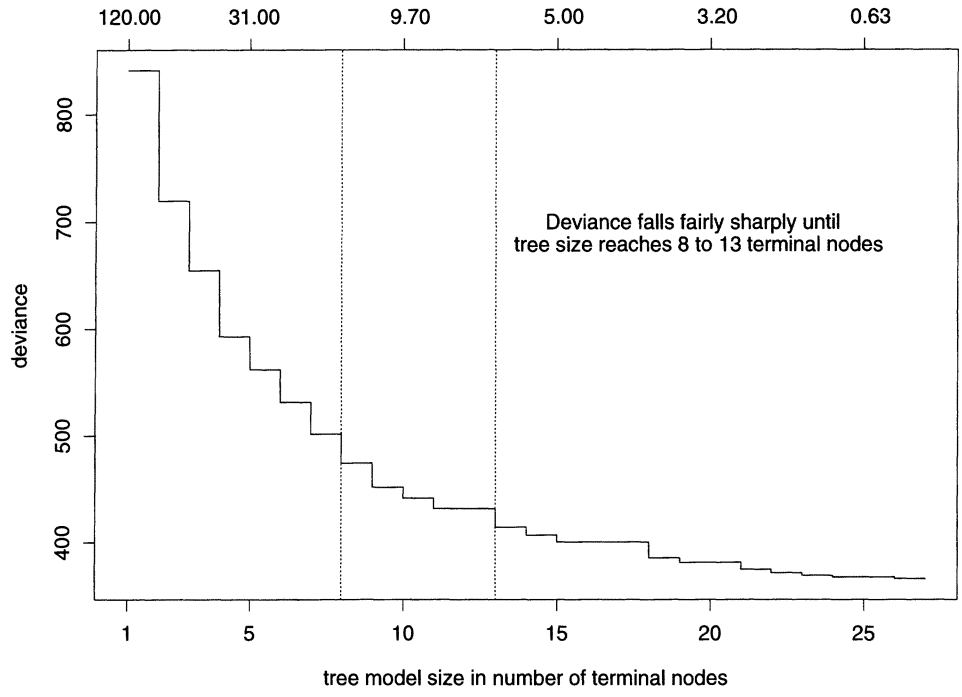

Figure 1. Shortleaf Pine: Deviance Curve for Increasing Tree Model Sizes. The two vertical lines indicate an area of the graph where the curve tends to flatten out, though this is a visual interpretation and therefore is somewhat arbitrary.

Denote the tree model from step 1 as T and let the deviance of this tree model be $D(T)$. Clark and Pregibon (1993) define a quantity $D_\delta(T)$, which we will call the penalized deviance for the model T, by

$$D_\delta(T) = D(T) + \delta \,\text{size}(T), \tag{2.2}$$

where δ is a penalty factor called a cost-complexity parameter and $size(T)$ is the number of terminal nodes in T. The quantity in (2.2) is analogous to Mallows C_p in the linear regression framework. For a given value of δ, the size of the tree model is determined in order to minimize $D_\delta(T)$. Divisions of the data that contribute least to a reduction in deviance are eliminated from the model. Note that there is a direct relationship between δ and the size of the model. A large value of δ will yield fewer terminal nodes. If $\delta = 0$, then there is no penalty for the model size and no pruning takes place. So δ is another tuning constant that determines how much a tree model is pruned. The question of interest is how large to make δ so that the tree model size is optimal in some sense. A value of δ was selected and, hence, the size of the model, through two techniques described below.

Technique 1: A Tree Deviance Curve. A tree deviance curve is a plot of the model deviance on the vertical axis versus the corresponding size of the model, which is shown on the lower horizontal axis (corresponding δ's for a particular model size are shown on the top horizontal axis). A plot of this type is shown in Figure 1. The deviance of a tree model

becomes smaller as the number of terminal nodes increases for the same reason that, in a linear model, R^2 becomes larger as more predictors are added to the model. The deviance curve shows how quickly the deviance drops as the model grows out of the root node. As the size of the model becomes larger, one can see that reductions in deviance become much less and that, eventually, very little reductions in deviance are gained by adding terminal nodes. From this type of plot, one can select the size of the model where the deviance curve tends to become horizontal. This is a subjective assessment, although there are criteria that can be used to quantify the value of adding additional terminal nodes to the model. For instance, if T is a model resulting from step 1 and T' is a model that has been pruned to a particular value of δ, then $1 - D(T')/D(G_0)$ is analogous to the R^2 criterion in linear regression.

Technique 2: Cross-Validation. We then perform an m-fold cross-validation as described in Clark and Pregibon (1993). Data are randomly separated into m subsets ($m \geq 2$) of approximately equal size, say n_S. One subset is held back or retained and a tree model is constructed using the data in the other $m - 1$ subsets. The tree model is pruned to a fixed value of δ, and the resulting model is used to predict the responses in the subset that were not used when building the model. The sum of the squared prediction errors is computed. This is done for each of the m subsets of the data for that fixed value of δ. The resulting squared prediction errors are all summed together and is called the cross-validation deviance. This process is repeated for a range of δ values that can be selected from the tree deviance curve, i.e., the range of δ should correspond to tree model sizes shown in the tree deviance curve. A plot of the cross-validation deviances (on the vertical axis) against δ (on the top horizontal axis) and $size(T)$ (on the lower horizontal axis) is called a cross-validation deviance curve. This plot is illustrated in Figure 2 and addressed further in Section 3.1. The plot will suggest a value of δ and, hence, it will suggest the size of a model with the minimum cross-validation deviance. Choosing m too small (i.e., 2 or 3) may result in a cross-validation curve being too erratic and difficult to interpret. A large value of m will produce a smoother curve but does require more computational time. Clark and Pregibon (1993) give an example using $m = 10$.

A cross-validation deviance curve depends on the random separation of data into m subsets. It is likely that two cross-validation deviance curves will be quite different for the same data set. Cross-validation deviance curves can be somewhat erratic due to the discontinuous nature of deviance for changing tree model size. For this reason, a large number, say B, of m-fold cross-validations on the same data set can be performed, and for each one, the model size with the minimum cross-validation deviance can be recorded. The result would be a distribution of model sizes where the minimum cross-validation deviance occurred. An example plot of minimum cross-validation deviances versus model size is shown in Figure 3, and it is discussed further in Section 3.1. One may decide to select a model size corresponding to the mode of the distribution of minimum deviances but, again, this selection is somewhat subjective. Only the model sizes corresponding to minimum cross-validation deviance are shown in Figure 3. It is possible that the cross-validation deviance is only slightly larger for larger models, and this possibility can be assessed by

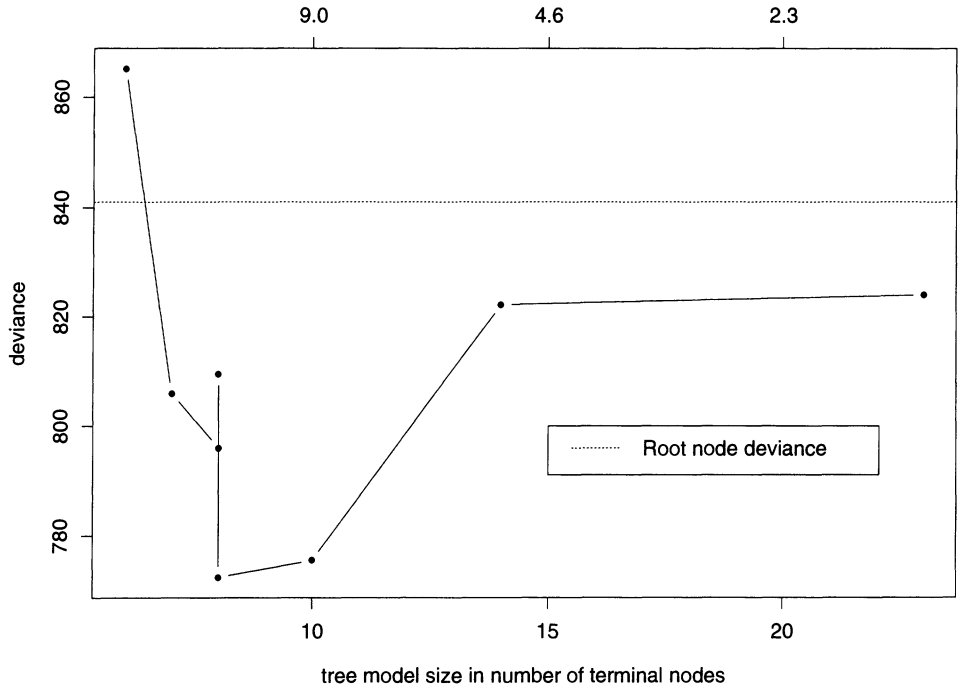

Figure 2. Shortleaf Pine: A Sixfold Cross-Validation With $\delta = 1, 5, 9, \ldots, 29$ and the Subsets Chosen So That the First Subset Is the First 29 Observations, the Second Subset the Second 29 Observations, and So on, With the Last Two Subsets Having 28 Observations Each. The root node deviance is the deviance of a one-node tree model (i.e., the sum of squares about the mean of all observations).

viewing several cross-validation plots like Figure 2. Since our objective was to use the model to capture the structure between the response and confounding variables before testing for the factor of interest, we often favored models with a size slightly larger than the size corresponding to the mode of the distribution of minimum deviances.

The choice of B is determined by the investigator, but it should be large enough to establish a pattern of optimal model sizes as determined by cross-validation. This process is computationally intensive, and large B can require substantial computing time, particularly if m is also large. The S-Plus function to perform this multiple m-fold cross-validation is available from the first author.

The results of the above steps allow the investigator to select a best size for the tree model. At this point, we have done nothing with the test factor, X_p. This factor will now be introduced within terminal nodes of the tree model. The objective is to determine if X_p is statistically significant for explaining any remaining pattern in the response values.

2.2 Testing the Significance of X_p

The test in which we are interested is analogous to a block design analysis of variance where grouped observations are blocks and the factor of interest is the treatment variable.

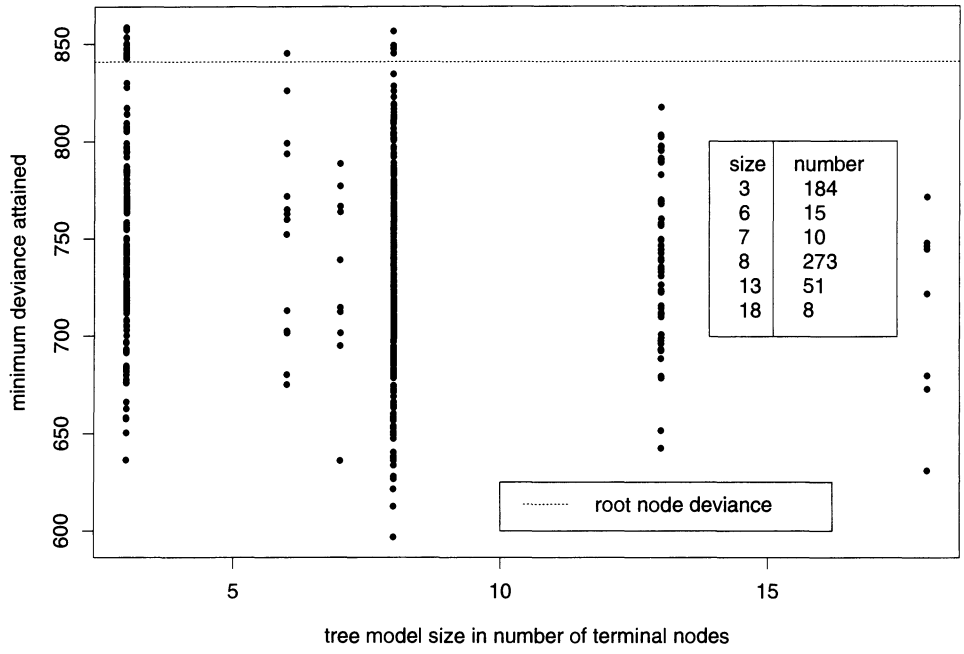

Figure 3. Shortleaf Pine: Minimum Deviance Versus Corresponding Tree Model Size for 500 Iterations of Sixfold Cross-Validations. The legend gives the number of minimum deviances occurring for a particular tree model size.

Parametric F-tests, allowing for unbalanced conditions, could be employed (Christensen 1987, Chapter 7). Alternatively, forms of nonparametric rank tests could be used (Lehmann 1975, p. 138).

We used a permutation test procedure reported in Mielke and Iyer (1982). This procedure requires no parametric assumptions, and it accommodates a lack of balance in terminal nodes. The procedure first aligns the blocks (i.e., terminal nodes) and then collapses them. It then performs the usual permutation test for a completely randomized design. The test statistic is compared with the reference randomization distribution under the null hypothesis. This is an approximate test. An exact test would require a randomization-based test accounting for block structure. However, this would require a Monte-Carlo approximation due to the large number of possible permutations, and this would be somewhat user dependent. With the procedure of Mielke and Iyer (1982), when the exact reference distribution becomes intractable due to a large data set, a Pearson type III distribution (matching of moments) is used to approximate the reference distribution (Mielke, Berry, and Brier 1981), thereby reducing the computational burden. The procedure reports a P-value for the null hypothesis that X_p has no effect on the response values once the confounding variables $(X_1, \ldots, X_{p-1})$ have been accounted for. The Fortran code to conduct this test is available from the first author (courtesy of Mielke and Iyer, referenced above). The tree-regression procedure and the permutation test are now illustrated in two examples.

3. EXAMPLES

3.1 A FORESTRY EXAMPLE

The first example data set was collected from the fourth cycle (spanning the years 1962–1972) and the fifth cycle (spanning the years 1972–1982) of USDA Forest Inventory and Analysis (FIA) measurements in Georgia. The objective of the analysis was to determine if there was a change in tree growth rates from the fourth cycle to the fifth cycle (hereafter referred to as a cycle effect). A sampled unit is a stand of trees, sometimes called a plot. The data are described in Table 1, and the data set is available from the first author. A more detailed description of the data is available in Bechtold et al. (1991). It is important to note that the data are not paired data. The two samples of stands from the fourth and fifth cycles were independent samples. There are various reasons for this that are discussed in Bechtold et al. (1991). Essentially, changes in the use of stands and destruction of stands from one 10-year period to the next made it very difficult to follow the same set of stands from one cycle to the next. Of the 172 observations in this example data set, only three stands reappeared in the fifth cycle, and since the structure of a stand will change dramatically from one cycle to the next, we believe it is of limited value matching these stands.

Bechtold et al. (1991) chose gross growth (GG) as their response variable, though net growth (NG) is an alternative candidate for a response variable. So Table 1 describes two available response variables (GG, NG) and five confounding variables (S, A, N, P, M). The mortality covariate (M) is part of the net growth response, NG. In an earlier report, the Bechtold et al. (1991) data sets were analyzed for three stand types (loblolly, slash, and shortleaf) and for both response variables (GG, NG) using the tree-regression procedure with the permutation test (Gadbury et al., 1998). We illustrate these procedures using the data set for shortleaf pine, and we chose net growth, NG, as our response variable. For the data analyzed, our tree regression model was

$$NG = f(A, N, S, P) + \epsilon. \tag{3.1}$$

Our first task was to group observations into subsets using model 3.1 and then to introduce the test factor for cycle effect, X_p, where

$$X_p = \begin{cases} 0 & \text{for observations in cycle 4} \\ 1 & \text{for observations in cycle 5.} \end{cases} \tag{3.2}$$

Table 1. Variable Descriptions of Bechtold et al. (1991) Data Sets

GG = gross annual basal area growth per acre (survivor growth + ingrowth)
M = annual basal area mortality per acre of trees $\geq$1 in. dbh (diameter breast height) alive at initial inventory that die from natural causes prior to terminal inventory
NG = net growth (*GG* − *M*)
S = site index representing volume growth potential (S represents 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

Tree-structured regression was performed on the data based on the model in (3.1), and the resulting tree regression model, before pruning, had 27 terminal nodes and a deviance of 365.6 (in squared units of the response NG). Default S-Plus values, $n_{\min} = 10$ and $\gamma = 0.01$, were used to construct this model. As stated before, models will be too large since the termination criteria of the tree regression algorithm are set to allow a large model. The tree deviance curve for this data set is shown in Figure 1. This curve suggests that only slight reductions in deviance are obtained after the tree model size increases beyond about 13 nodes, though this is a subjective visual assessment.

The 27-node model was then pruned. An $m = 6$-fold cross-validation was used so that there were about $n_S = 29$ observations in each of six subsets. Figure 2 shows a sample of one cross-validation deviance curve. Several such plots were viewed to determine how much the tree model cross-validation deviance fluctuated for changing tree model sizes. The particular plot shown in Figure 2 seems to suggest an 8–10-node model, though this is only one cross-validation result.

A plot of the distribution of the optimal tree model sizes, as determined by cross-validation deviance, for each of $B = 500$ cross-validations is shown in Figure 3. The mode of the distribution corresponds to an eight-node tree model. However, there are 51 cases where the minimum cross-validation deviance occurred for a 13-node tree model. This is roughly 10% of B, and we tended to favor this larger model since we prefer to accept a slight overfit of the model to the data for reasons discussed earlier. The numbers in the legend of Figure 3 sum to 541. For some cross-validations, two values of δ could have corresponded to the same tree model size. If this model size corresponded to the minimum deviance for that cross-validation, it would have been counted twice in the S-Plus function that produced the plot.

The tree model we chose for this data set is shown in Figure 4. The resulting deviance for the 13-node model was 414 (squared NG units). At this point, the cycle effect was tested within terminal nodes of the tree model using the permutation procedure. In the procedure, the 13 subsets were aligned, then collapsed, and a permutation test for a cycle effect was performed. The reported P-value was for a two-sided alternative against the null hypothesis that there was no cycle effect. The P-value was equal to 0.012, suggesting that there was evidence in the data that stand net growth rates changed from one cycle to the next after effects of other covariates had been taken into account.

As mentioned earlier, Bechtold et al. (1991) did not analyze the net growth response variable, but Ueng, Gadbury, and Schreuder (1997) analyzed net growth for this same data set using a robust linear model-fitting procedure to determine a best model. For comparison, they reported a P-value of 0.0051 for the same hypothesis and for this same data set.

There were no required assumptions regarding the distributional form of these data. Another advantage is that no form of the model was assumed a priori. The tree model algorithm suggested a form based on data values. Higher order terms and interactions were automatically accommodated when the tree model included multiple divisions of the data based on the same covariate. Furthermore, the tree model was invariant to monotonic transformations of covariates, and outliers only affected the particular node in which they

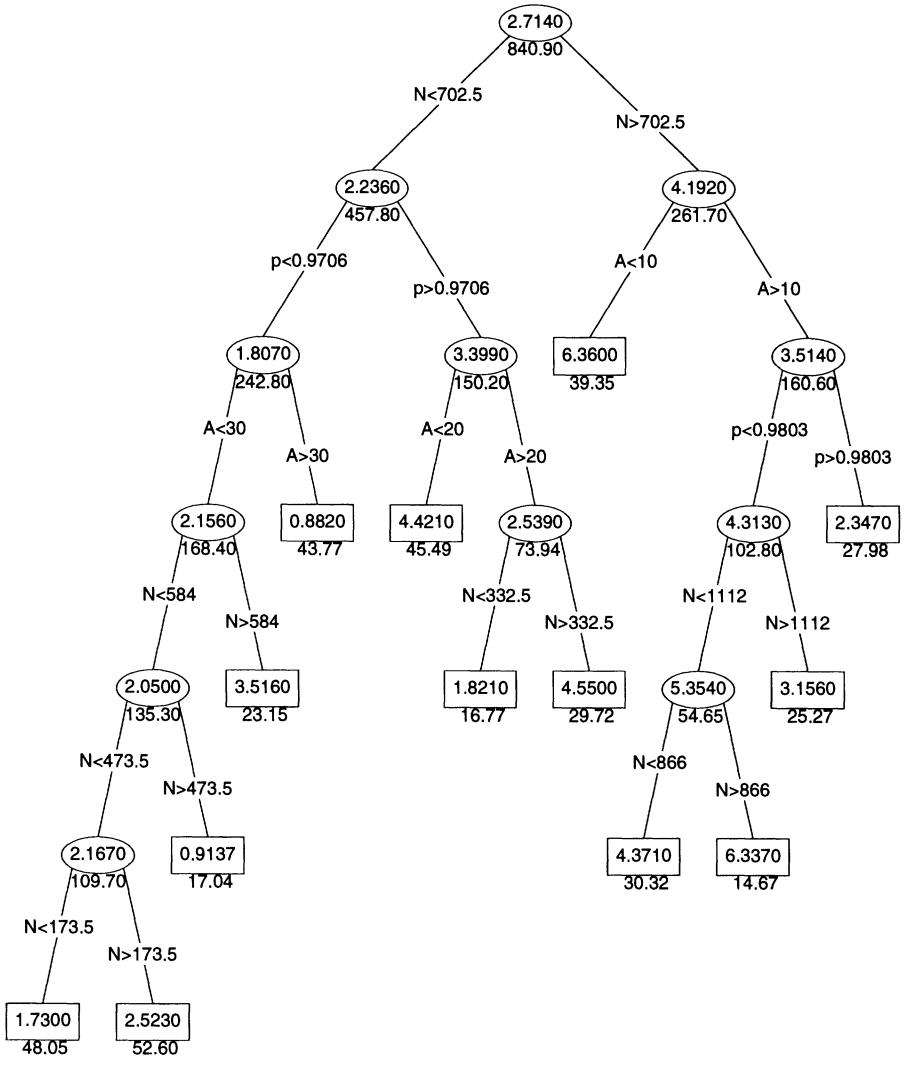

Figure 4. Shortleaf Pine: A 13-Node Tree Model. Rectangles are terminal nodes. The number inside the rectangle/oval is the predicted value (mean value) at that node. The number below the rectangles/ovals is the deviance at that node.

resided, as opposed to a linear model, where they may affect the entire model. For these reasons, we feel that this procedure is a useful technique to perform analysis of covariance on large survey data sets like this one where relationships among variables appear to be complex.

The adaptive methodology described in this article is clearly an approximate inference procedure in the sense that the P-values reported by the analysis are not expected to be exact. However, for the procedure to be reliable for our application, tests of a cycle effect on growth rates conducted using a nominal α level should have an actual α level close to the nominal value. In order to assess this closeness, we conducted a small simulation study,

reported in Gadbury et al. (1998). The FIA data sets were used as models to generate simulated data. The simulation results provided some evidence that reported P-values using these techniques were close to actual P-values even when there was a fairly high correlation between cycle and the covariates.

3.2 AIR POLLUTION EXAMPLE

McDonald and Schwing (1973) analyzed a data set to assess the effect of air pollutants on age-adjusted mortality. There are 16 variables and 60 observations in this data set, and it is available at the internet site `http://lib.stat.cmu.edu/datasets`. One can see McDonald and Schwing (1973) for a more detailed description of the data. Here we reference only those variables relevant to this analysis. The response variable is $MORT$, which is the total age-adjusted mortality rate expressed as deaths per 100,000 population. The interest is in determining if certain pollutants are linked to increased mortality. There are three pollutant variables: the relative pollution potential of hydrocarbons (HC), oxides of nitrogen (NO_x), and sulfur dioxide (SO_2). The 12 confounding variables relate to various weather and socioeconomic measures in each of the 60 standard metropolitan statistical areas (SMSA) where the data were obtained.

McDonald and Schwing (1973) found no association between $MORT$ and the two pollutants HC and NO_x. This agreed with our analyses using the methods in this article. McDonald and Schwing (1973) found a substantial positive association between SO_2 and $MORT$. They used a variable selection method using the ridge trace, and it resulted in the model

$$\begin{aligned} MORT = {} & 988.4 + 1.487PREC + 1.633JANT - 11.533EDUC \\ & + 0.004DENS + 4.145NONW + 0.245SO_2 + \epsilon, \end{aligned} \tag{3.3}$$

where $PREC$ is mean annual precipitation (inches), $JANT$ is mean January temperature (degrees F), $EDUC$ is median school years completed for those over 25 in 1960, $DENS$ is population per square mile in urbanized area in 1960, and $NONW$ is percent nonwhite population in urbanized area in 1960.

We fit a tree model using $MORT$ as the response variable and all 12 confounding variables as covariates (i.e., we did not include any pollutant variables in the model-building step). Since the number of observations was only 60 and since the root node deviance was large (228,300 in squared $MORT$ units), we set $n_{\min} = 6$ and $\gamma = 0.0001$ in step 1 of the model-building procedure. The result was a 16-node tree model with a deviance of 23,820. Using the tree deviance curve and the cross-validation procedure discussed earlier, we selected a nine-node tree model that we designate as T_9. Incidentally, cross-validation deviance curves tended to show a low cross-validation deviance (i.e., close to the minimum) for models even larger than nine nodes, but nine nodes seemed sufficient for our analysis. The deviance for T_9 is 38,310 and the value of $1 - D(T_9)/D(G_0)$ is 0.83. As a comparison, the value of R^2 for a linear regression model fitting $MORT$ to all 12 confounding variables is 0.72.

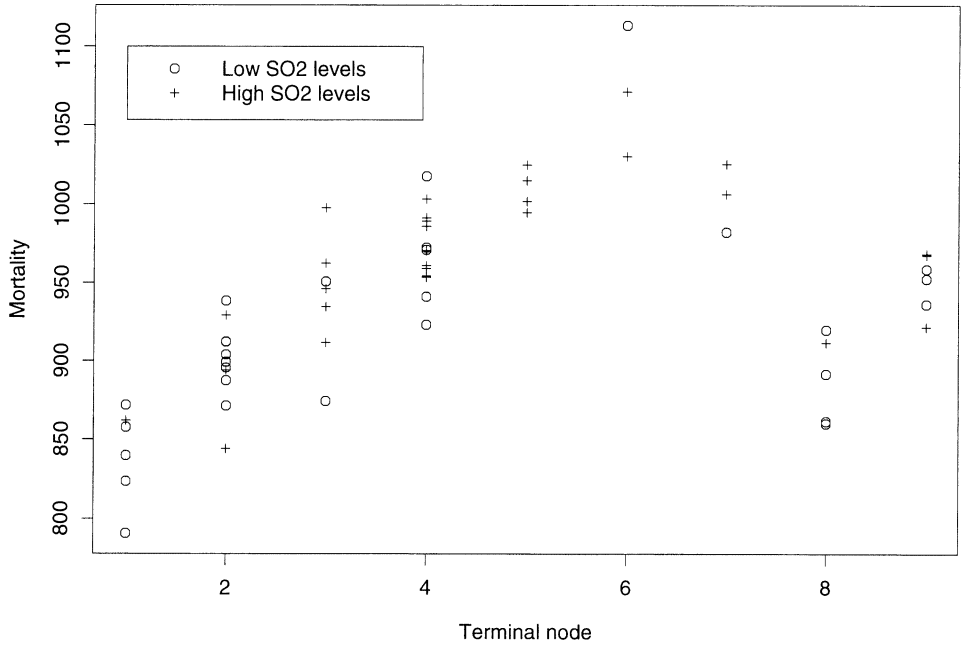

Figure 5. Air Pollution Data. Mortality values within each of nine terminal nodes of a tree regression model. High and low values of SO_2 are determined by whether an observation is above or below the median SO_2 value.

Our selected model, T_9, used five confounding variables, four of which are included in the model given by Equation (3.3). Our model did not use $JANT$ but did use the variable $POPN$, which is the population per household, 1960. The structure of the model suggested that there are interactions among confounding variables. For example, an interaction between $NONW$ and $EDUC$ was apparent.

We created a dichotomous test factor by collapsing SO_2 into two categories determined by whether an observation's SO_2 levels were above or below M_{SO_2}, the median of the 60 SO_2 values, i.e., the value of the test factor for an observation i, X_{pi}, $i = 1, \ldots, 60$, was determined by

$$X_{pi} = \begin{cases} 0 & \text{if } SO_{2\,i} < M_{SO_2} \\ 1 & \text{if } SO_{2\,i} \geq M_{SO_2}. \end{cases} \tag{3.4}$$

The permutation procedure then tested for the effect of X_p on $MORT$ within terminal nodes of T_9. The P-value from the test was 0.348, providing no evidence of an association between SO_2 levels and $MORT$ after adjusting for confounding variables.

Figure 5 shows a plot of the mortality values in each terminal node of T_9. Two plotting characters were used to identify observations with high values of SO_2 (i.e., with $X_p = 1$) and those with low values of SO_2 ($X_p = 0$). The plot does not reveal any visual systematic pattern within terminal nodes between high and low SO_2 values and corresponding mortality values.

An information loss occurred when the continuous variable SO_2 was collapsed into

two levels. For this reason, we conducted an alternative test within terminal nodes of T_9. Let $I_{i,j}$ be an indicator variable that is equal to one if observation i is in terminal node j, $i = 1, \ldots, 60$ and $j = 1, \ldots, 9$. We fit the linear model

$$MORT_i = \alpha_j \, I_{i,j} + \beta \, SO_{2\,i} + \epsilon_i;$$

i.e., the terminal node of the tree model was a blocking variable and we assumed equal slopes of the regression model relating $MORT$ and SO_2 within terminal nodes. Diagnostics did not indicate any serious departures from normality assumptions. The P-value of the test, H_0: $\beta = 0$, versus a two-sided alternative was 0.101, which indicated only marginal evidence of an association between $MORT$ and SO_2.

Finally, we aligned the terminal nodes of T_9 by subtracting the mean response within each terminal node from the observations in that node. We then performed a linear regression of the aligned $MORT$ responses on SO_2. Again, no significant association between $MORT$ and SO_2 was detected (P-value = 0.136).

The structure in this data set appears to be complex and our approach is particularly suited to such data. But one might argue that our approach is conservative since we only use confounding variables in the model-building step before considering SO_2. This was intentional because we feel that an apparent association between SO_2 and mortality will often be perceived as a cause–effect relationship.

We conclude that these data do not provide convincing evidence of a link between SO_2 and mortality once confounding variables have been taken into account. This does not mean that there is no such link. One might need to resort to other criteria outside these data to strengthen the claim that there is a causal connection between high SO_2 levels and mortality. For instance, see Hill (1965) and/or Susser (1988).

4. DISCUSSION

We have described an analysis of covariance procedure that requires very minimal assumptions about the data. Tree regression was used to adjust for possible confounding effects of covariates, and a permutation procedure was used to test for the effect of a factor of interest, conditional on the covariates. If the data were paired data, the procedure could be adapted by altering the permutation test to one that randomizes the sign on observations in terminal nodes.

The procedure was illustrated in two examples. In the forestry example, there remained a significant change in tree growth rates after adjusting for the covariates. In the air pollution example, the effect of SO_2 on mortality was confounded with effects of weather variables and socioeconomic variables. There was little remaining association between SO_2 and mortality after adjustment for covariates.

A limitation of the standard tree regression technique is that the covariate space is divided into hyperrectangles by the standard procedure of dividing data into subsets. This may not be effective if, e.g., the relationship between the response and covariates is highly linear. Consider an extreme case, $Y = X$. In this case, the standard tree growing algorithm

will divide data based on the same covariate X multiple times while trying to capture the structure between Y and X. To address this possibility, some have suggested allowing the data to divide on linear combinations of the variables (cf., Breiman et al. 1984). In more complicated examples, determining a best linear combination of variables on which to divide the data can be difficult. Breiman et al. (1984) described an algorithm that searches for such a linear combination. Drawbacks are increased computation time, loss of interpretability, higher cross-validated errors, and the fact that the tree model is no longer invariant to monotonic transformations of predictor variables (Breiman et al. 1984). We felt that, for our application, the advantages of standard tree regression offset the limitation.

The standard procedure uses a reduction in sums of squares about the mean as a criteria to determine where to divide the data into subsets. Others have suggested least absolute deviations (LAD) as an alternative criteria (Breiman et al. 1984) that would likely be more robust when there are extreme outlying observations in the response variable.

In summary, the method reported in this article appears to be a useful, robust (not restricted to model assumptions that include both assumed model forms and assumed distributional forms of the data), and interpretable approach to examine whether a factor of interest (e.g., a treatment) affects a response or whether change has occurred over time in variables of interest using large-scale survey data sets when confounding variables are present.

ACKNOWLEDGMENTS

We thank the editor, Dr. Bryan Manly, for helpful comments that improved the manuscript.

[Received August 1999. Accepted February 2001.]

REFERENCES

Bechtold, W. A., Ruark, G. A., and Lloyd, F. T. (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 Stone, C. J. (1984), *Classification and Regression Trees,* Monterey, CA: Wadsworth.

Christensen, R. (1987), *Plane Answers to Complex Questions,* New York: Springer-Verlag.

Clark, L. A., and Pregibon, D. (1993), "Tree-Based Models," in *Statistical Models in S,* eds. J. M. Chambers and T. J. Hastie, New York: Chapman and Hall, pp. 377–419.

Gadbury, G. L., Iyer, H. K., Schreuder, H. T., and Ueng, C. Y. (1998), "A Nonparametric Analysis of Plot Basal Area Growth in Southeastern USA Using Tree Based Models," Research Paper RMRS-RP-2, USDA Forest Service Rocky Mountain Forest and Experiments Station.

Hill, A. B. (1965), "The Environment and Disease: Association or Causation?" *Proceedings of the Royal Society of Medicine,* 58, 295–300.

Lehmann, E. L. (1975), *Nonparametrics: Statistical Methods Based on Ranks,* San Francisco: Holden-Day.

McDonald, G. C., and Schwing, R. C. (1973), "Instabilities of Regression Estimates Relating Air Pollution to Mortality," *Technometrics,* 15, 463–481.

Mielke, P. W., Berry, K. J., and Brier, G. W. (1981), "Application of Multi-Response Permutation Procedures for Examining Seasonal Changes in Monthly Sea-Level Pressure Patterns," *Monthly Weather Review,* 109, 120–126.

Mielke, P. W., and Iyer, H. K. (1982), "Permutation Techniques for Analyzing Multi-Response Data From Randomized Block Experiments," *Communications in Statistics—Theory and Methods,* 11, 1427–1437.

Rosenbaum, P. R. (1995), *Observational Studies,* New York: Springer-Verlag.

Susser, M. (1988), "Falsification, Verification, and Causal Inference in Epidemiology: Reconsiderations in the Light of Sir Karl Popper's Philosophy," In *Causal Inference,* K. J. Rothman, ed. Chestnut Hill, MA: Epidemiology Resources.

Ueng, C. Y., Gadbury, G. L., and Schreuder, H. T. (1997), "Robust Regression Analysis of Growth in Basal Area of Natural Pine Stands in Georgia and Alabama for 1962–1972 and 1972–1982," Research Paper RM-RP-331, USDA Forest Service RM Forest and Experiments Station.