USING TREE REGRESSION TO IDENTIFY NUTRITIONAL AND ENVIRONMENTAL FACTORS AFFECTING SUGARCANE PRODUCTION

Kenneth M. Portier

David L. Anderson

USING TREE REGRESSION TO IDENTIFY NUTRITIONAL AND ENVIRONMENTAL FACTORS AFFECTING SUGARCANE PRODUCTION

Kenneth M. Portier[1] and David L. Anderson[2]
[1]Department of Statistics, Institute of Food and Agricultural Sciences
University of Florida, Gainesville, FL 32611-0560
[2]Everglades Research and Education Center, Institute of Food and Agricultural Sciences
University of Florida, Belle Glade, FL 33430-8003

ABSTRACT

A prediction function is developed for sugarcane yield using preplant soil nutrition levels, cultivar, and soil type. A tree regression approach is used because the resulting function encompasses the complexity of response between yield, multiple nutrients and other factors, while handling large amounts of data and providing information useful in the development of fertilizer and other production recommendations. Data collected from 148 control plots of experiments performed on commercial fields in the Everglades Agricultural Area of Florida are used to illustrate the method.

Keywords: Binary tree, Soil testing, Florida Everglades Agricultural Area

1.0 Introduction

It is generally accepted that optimal plant growth and yield is conditional on plants having access to adequate nutrition. Too little of any one nutrient limits growth, regardless of the levels of other nutrients. As the level of a limiting nutrient is increased, yields increase up to the point that another nutrient is limiting or the plant has reached its full genetic potential. In a controlled yield trial these effects can be demonstrated, but usually by fixing the levels of most nutrients and varying the levels of only a couple nutrients.

In this study, a predictive equation is developed for crop yield using preplant soil nutrition and other factors affecting sugarcane grown in South Florida. The objective was to relate levels of nutrients in common soil test analysis and generally known crop and location characteristics, such as cultivar type, soil type and 'crop' year, to sugarcane yield. Such a predictive function would be very useful to the soil test lab in making fertilizer recommendations.

Because the data are not from controlled experiments and because of the effects discussed above, standard linear model analysis of these data provide predictions with very low precision. Other dimension reduction techniques, such as principle component regression, do not improve significantly over the original glm's. The inclusion of

interaction effects into the model increases the degree of multicollinearity in the model but does nothing to improve prediction.

Among the general class of robust regression techniques, tree regression offers the best chance of developing a prediction function which encompasses the complexity of response while providing a function which could easily be used by the soil test lab. Because tree regression is still a relatively new technique and not found in general statistical methodology text books, a review is provided in the next section.

2.0 Tree Regression

Let the real-valued variable y denote the response or dependent variable and $X = \{X_1 \ldots X_p\}$ denote a vector of independent or predictor variables from the measurement space Ω. In a regression analysis, our objective is either to 1) explore the structure of the relationship between y and X or 2) predict y for future values of X. This goal is usually accomplished through the development and estimation of a real-valued predictor function, denoted $f(x)$ defined on Ω.

If the predictor function is known except for a finite set of parameters θ, that is $E(Y|X = x) = f(x, \theta)$, least square techniques can be used to estimate θ. If further the form of the function is assumed linear in these parameters, elegant estimation and goodness-of-fit procedures are available.

What if the form of the predictor is not known, and the dimensionality of X is such that some form of variable selection is required? Assuming a linear predictor, either a stepwise or best subsets selection algorithm is usually recommended. The properties of these procedures and associated diagnostic tools are generally known. Tree-based regression, like other robust regression techniques, facilitates development of predictors when its general form is not known.

In tree-based regression, the measurement space, Ω is partitioned by repeated binary splitting of subsets of Ω into two descendent subsets, starting with Ω itself. The partition of Ω is usually represented as a binary tree (Figure 1), denoted T, with nodes, t_k, $k = 1, 2, \ldots, K$. This binary recursive partitioning of Ω is done in such a way that the within node variability in observed response is minimized. That is, the added residual variance of the two descendent nodes is less than the residual variance in the original ancestor.

If we let Ω_k be that subset of Ω represented in the k^{th} node, then the value of the predictor function for node k is the mean of the observed responses in that node, that is $f(x|x \in \Omega_k) = \bar{y}(x) = \bar{y}(t_k)$. In this way, the regression surface is approximated by a step function (see Figure 2).

The residual sums of squares (or deviance) associated with $f(\boldsymbol{x})$, for the sample $(y_i, \boldsymbol{x}_i)$ $i = 1, 2, ..., n$ is

$$R(f) = \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2 \tag{1}$$

For node t_j of tree T, define the associated residual sums of squares or residual deviance as:

$$r(t_j) = \sum_{i=1}^{n} I(\boldsymbol{x}_i \in t_j)(y_i - \bar{y}(t_j))^2 \tag{2}$$

where $I(\bullet)$ is the indicator function. The residual sums of squares for tree T is the sum of all terminal node-associated sums of squares. That is, if t_j^*, $j = 1, 2, ..., k^*$ are the terminal nodes of tree T (see Figure 1), then

$$R(T) = \sum_{j=1}^{k^*} \sum_{i=1}^{n} I(\boldsymbol{x}_i \in t_j^*)(y_i - \bar{y}(t_j^*))^2 \tag{3}$$

The measure of improvement observed by splitting node t_j into nodes $t_{j'}$ and $t_{j''}$ is given by

$$r(t_j) - r(t_{j'}) - r(t_{j''}). \tag{4}$$

Using this measure of improvement, the best split at a node is that split which most successfully separates the high response values from the low ones. That is, for a successful split, $\bar{y}(t_{j'}) < \bar{y}(t_j) < \bar{y}(t_{j''})$ and $R(unsplit\ \ tree) > R(split\ \ tree)$.

Candidates for splits are defined through the individual components of $\boldsymbol{X}$. Let $\boldsymbol{X}$ be composed of p components, that is the i^{th} observation is $\boldsymbol{x}_i' = \{x_{i1}, x_{i2}, ..., x_{ip}\}$. For an ordinal, interval or ratio component, say X_ℓ, splits can be created as

$$\begin{aligned} t_{j'} &= \{y_i \mid x_{\ell i} < x^0\} \\ t_{j''} &= \{y_i \mid x_{\ell i} > x^0\} \end{aligned} \tag{5}$$

for each observed value of x^0 in the domain of X_ℓ. Similarly, if X_ℓ is a nominal scale variable with observed levels $C_\ell = \{C_{\ell 1}, C_{\ell 2}, ..., C_{\ell L}\}$, then splits are created as

$$t_{j'} = \{y_i \mid x_{\ell i} \in C',\ C' \subset C\}$$
$$t_{j''} = \{y_i \mid x_{\ell i} \in \bar{C}',\ \bar{C}' \subset C,\ \bar{C}' \cap C' = \phi,\ \bar{C}' \cup C' = C\} \quad (6)$$

For each subset Ω_k, the split, which maximizes the improvement in residual sums of squares is chosen from among all possible splits across all components. In this way, each component is examined at each opportunity to split. This partitioning of Ω is usually performed until either 1) each terminal node has fewer than a prespecified number of observations (usually 5) or 2) the variability in each terminal node is less than some small value, usually defined to be 1% of the variance of the root node. This results in a tree which is much larger than the data can support. As with over-specification in multiple linear regression, the final predictor function is over-optimistic in that the estimate of residual variance is much smaller than it should be.

To correct for over-optimistic trees, the maximum-sized tree is pruned back to a smaller tree. Our objective is to simplify the tree without significantly sacrificing goodness-of-fit. The pruning process produces a nested sequence of trees determined by recursively snipping off the least important branches. Importance is measured by a cost-complexity index

$$R_\alpha = R(T) + \alpha |T| \quad (7)$$

where $|T|$ = number of nodes in tree T, and α is the cost-complexity parameter. For a given value of α, the optimal subtree minimizes the cost-complexity index. Optimal subtree residual variance decreases and tree size increases monotonically as α decreases. Venables and Ripley (1994) use Mallow's C_p statistic to suggest using $\alpha = K\hat{\sigma}^2$ where $\hat{\sigma}^2$ is the residual mean square for the maximum size tree and K is a value between 2 and 6.

Supplemental information on optimal tree size is obtained by examination of tree prediction error using a 10-fold cross-validation process. Here the data are divided into 10 roughly equal parts and a tree is developed using 9 parts with prediction error (residual variances) computed on the tenth part. This is done in the 10 possible ways, with prediction error averaged for optimal subtrees for specified values of α. A plot of prediction error by optimal subtree size is not necessarily monotonically decreasing but trees with minimum prediction error are candidates for optimal tree.

3.0 Data

The Florida Everglades Agricultural Area (EAA) supports the production on 155,000 ha of organic (muck) soils and 27,000 ha of mineral (sand) soil. From 1983 through 1989, 14 major yield trials were held on commercial lands in the EAA. For this study soil test information was collected from the 148 untreated (control) plots associated with these

studies (Korndörfer et al 1995). Plot soils were sampled one month after initial planting and at the beginning of each ratoon crop. A ratoon crop is the new sugarcane produced from the root mass or stool which remain in the ground after the previous years growth has been cut off. Up to three ratoon crops may be grown on the same roots.

The predictor variables for this study were:

Continuous Variables

K	level of potassium in the soil
pH	pH level of the soil
Mg	level of soil magnesium
Pw	level of phosphorus extracted from the soil using water
Pa	level of phosphorus extracted from the soil using a dilute acid

Classification Variables

Sname	soil type name (Pahokee, Terra Ceia, Tonng, Lauderhill and Okeelanta)
Cult	variety id code (61620, 681026, 701133, and 742004)
Crop	1 = 1st year plant growth 2 = 1st ratoon growth 3 = 2nd ratoon growth 4 = 3rd ratoon growth

The response variable was sugarcane yield for the plot measured in tons cane per acre, denoted **TPA**.

4.0 Analysis

Data were input and managed using SAS (SAS Institute Inc. 1991). Tree regression analysis computations were performed using S-Plus (Clark and Pregibon, 1991; Venables and Ripley, 1994, Breiman et al 1984).

Site yields (**TPA**) were first modified to remove year-to-year variability in overall yield by computing the deviation of site yield to average annual yield, and adding this to the yield averaged over all sites and years. In this way, year was treated as an extraneous source of variation and removed prior to model building.

With only 148 observations, the full tree is not very large. Using the stopping rules discussed in section 2, the full tree (Figure 3) has 21 terminal nodes and explains approximately 81% of total variability in the year mean adjusted yield. Starting at the top of the diagram in Figure 3, the full 148 plots are divided based on what the observed value of **Pa** was less than 13.076 ppm (to the left branch) or greater than or equal to 13.076 ppm (to the right branch). The overall mean value for **TPA** is 48.5 tons. After the split, the observations in the left branch have mean 39.2 tons and the right branch has mean 54.8. The value in parenthesis is the deviance of the node, hence the initial tree has deviance 22520 whereas, after the split, the left node deviance is 3671 and the right node deviance is 10110. Thus the improvement for this split is 8739 from (eq. 4). The residual mean deviance for the maximum tree is 28.04.

The quality of this fit is illustrated by a plot of observed **TPA** versus predicted **TPA** just as one would do in regression analysis. The discreteness of the prediction function is evident in Figure 4. A plot of residuals versus predicted values, Figure 5, indicates no obvious outliers although there are a couple of points which are not predicted very well.

Tree deviance as a function of optimal tree size for specified values of the complexity parameter is plotted in Figure 6. Using the suggestion of Venables and Ripley and selecting $K = 6$ we have the optimal α computed as $6\hat{\sigma}^2 = 6 \times 28.04 = 168.24$. This translates into an optimal tree size of 10 nodes. This suggestion of 10 nodes in the optimal tree is further supported by the cross-validation deviance plot, Figure 7, which shows a minimum deviance at 10 nodes.

The best 10-node tree is given in Figure 8. The residual mean deviance for this tree is 35.42, and the fraction of total variability explained is 75.4%. Plots of observed versus predicted (Figure 9) and residual versus predicted (Figure 10) are provided for comparison with similar graphs (Figures 4 and 5) for the full tree.

5.0 Discussion and Summary

Variables which are used in the initial tree splits usually have the greatest impact on the response. In this case, the acid extracted soil phosphorus level was important as a predictor of yield (**TPA**). Further, it seems that in low phosphorus sites, crops beyond the first ratoon (**Crop** = 3 or 4) show lower yields than do earlier crops. For higher phosphorus areas, soil potassium are important, but the yield differences are not as large. Further examination of the tree suggest situations where yields from low levels of one soil nutrient may be increased by higher levels of another nutrient.

The tree shown in Figure 8 presents a predictor function quite unlike what would be expected from a general linear model. The tree shows complex interactions and variable relationships which would be difficult to observe from a multiple regression. To find the predicted value for a new observation, the observation is "dropped" down the tree until it

settles into a node. The mean value for that node is the predicted value for the observation. If the new observation has a missing value for a predictor variable in the decision path, the predicted value becomes the mean of the last node attained.

Results from a tree regression are difficult to extrapolate beyond the domain of the existing data. If the underlying relationships are truly continuous and polynomial in the quantitative predictor variables and the effects of qualitative factors are global, tree regression will produce an over-parameterized model which may not predict as well as an appropriate polynomial.

The utility of tree regression methods in the analysis of crop nutrition/yield data is suggested by this study. For example, the final model suggests (assuming the biology minimally supports cause and effect, as it does in this case) that by getting the acid-based soil phosphorus level above 13.076 ppm one might expect a much better yield. In addition, if soil **K** were low (less than 54.5 ppm), getting soil **P** above 38.5 ppm might be expected to provide an additional yield increase. The cost of amending the soil to reach these levels can be compared to the expected benefits in increased sugarcane yield. Similarly, cultivar changes and whether to replant after year 2 can all be examined through this model.

It is still too early to know how this information can be used in developing recommendations from a soil testing lab. Traditionally, recommendations have been based on a decision rule which looks at the nutrients separately. The tree regression results indicate that the decision to supplement one nutrient (say **Mg**) might depend on the levels of other nutrients (**Pa** and **K**). Multivariate decision rules of this type have the potential to save growers the cost of amendments in situations where the full benefit of the amendment may not be realized in increased yield.

6.0 Acknowledgements

We wish to thank the following who assisted and cooperated in the field experiments from which these data were collected: M. F. Ulloa, New Hope Sugar Corp., Pahokee, FL; W. Browning and G. Crews, A. Duda and Sons, Belle Glade, FL; A. Sanchez, Camayen Cattle Corp., Pahokee, FL; M. Perro, Okeelanta Corp., South Bay, FL; and K. Shuler, University of Florida, Coop. Ext. Service, Palm Beach, FL. We also thank Dr. G. H. Korndorfer, Federal University of Uberlandia, MG, Brazil, for compiling the study data while a visiting professor at the Everglades Research and Education Center, Belle Glade, FL. Contribution from the University of Florida, Institute of Food and Agricultural Sciences, Florida Agricultural Experiment Station, Journal Series.

7.0 References

Breiman, L. J.H. Friedman, R.A. Olsen and C.J. Stone, *Classification and Regression Trees*, Wadsworth International Group, Belmont, CA, 1984.

Clark, L.A. and D. Pregibon, "Tree-Based Models" in *Statistical Models in S*, Chambers, J.M. and T.J. Hastie (eds), Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1991, 377-419.

Korndörfer, G.H., D.L. Anderson, K.M. Portier, and E.A. Hanlon, 1995, "Phosphorus soil test correlation to sugarcane grown on histosoils in the Everglades," *Soil Sci. Soc. Am. J.* 59:1655-1661.

SAS Institute Inc., *SAS User's Guide: Language and Procedures, Usage 2, Version 6, First Edition*," Cary, NC:SAS Institute Inc., 1991, 649 pp.

Venables, W.N. and B.D. Ripley, 1994. *Modern applied statistics with S-PLUS*, Springer-Verlag New York, Inc., New York, NY, Chapter 13.

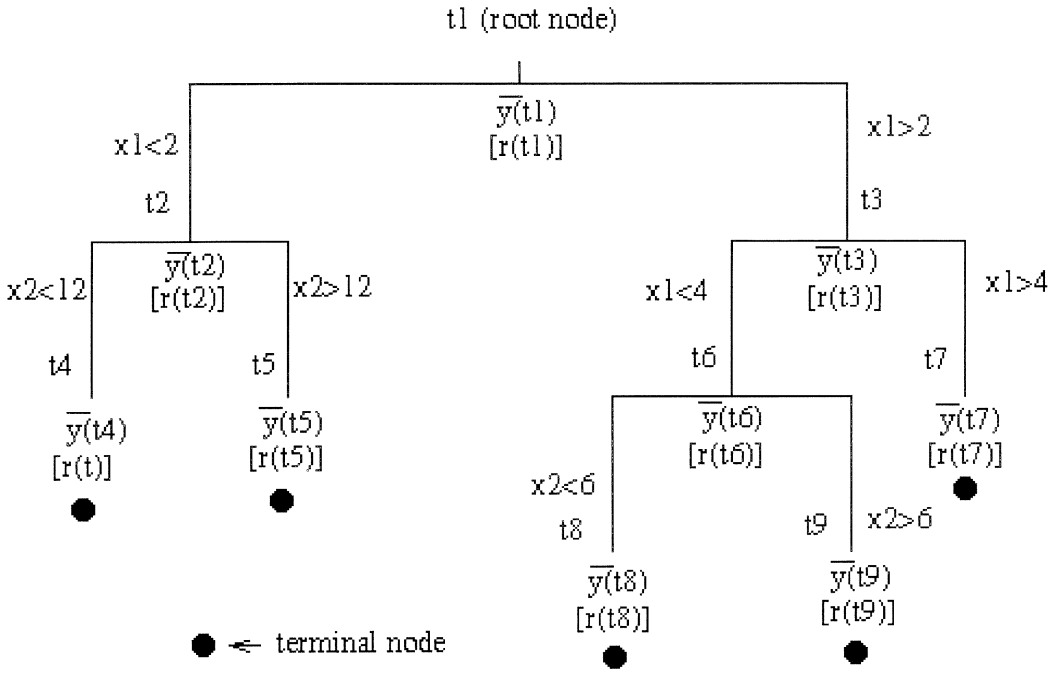

Figure 1. Example of a binary tree.

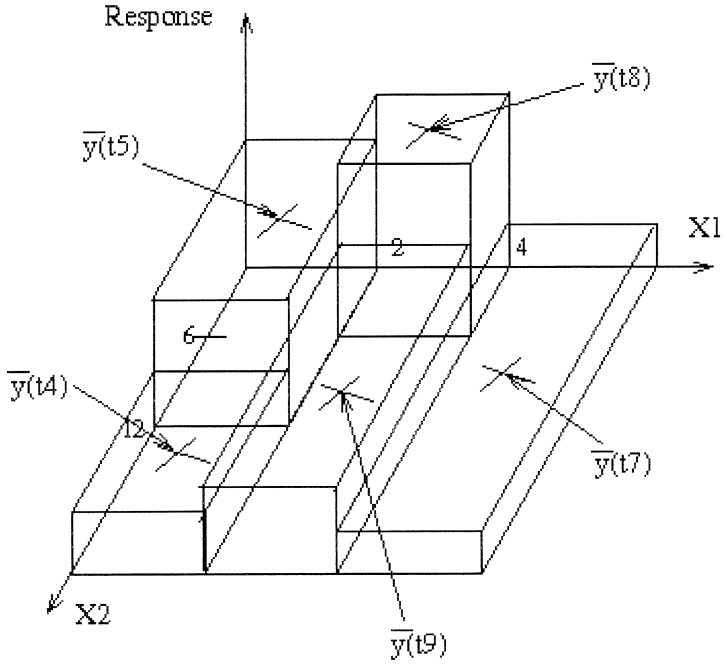

Figure 2. Regression surface approximated by a step function.

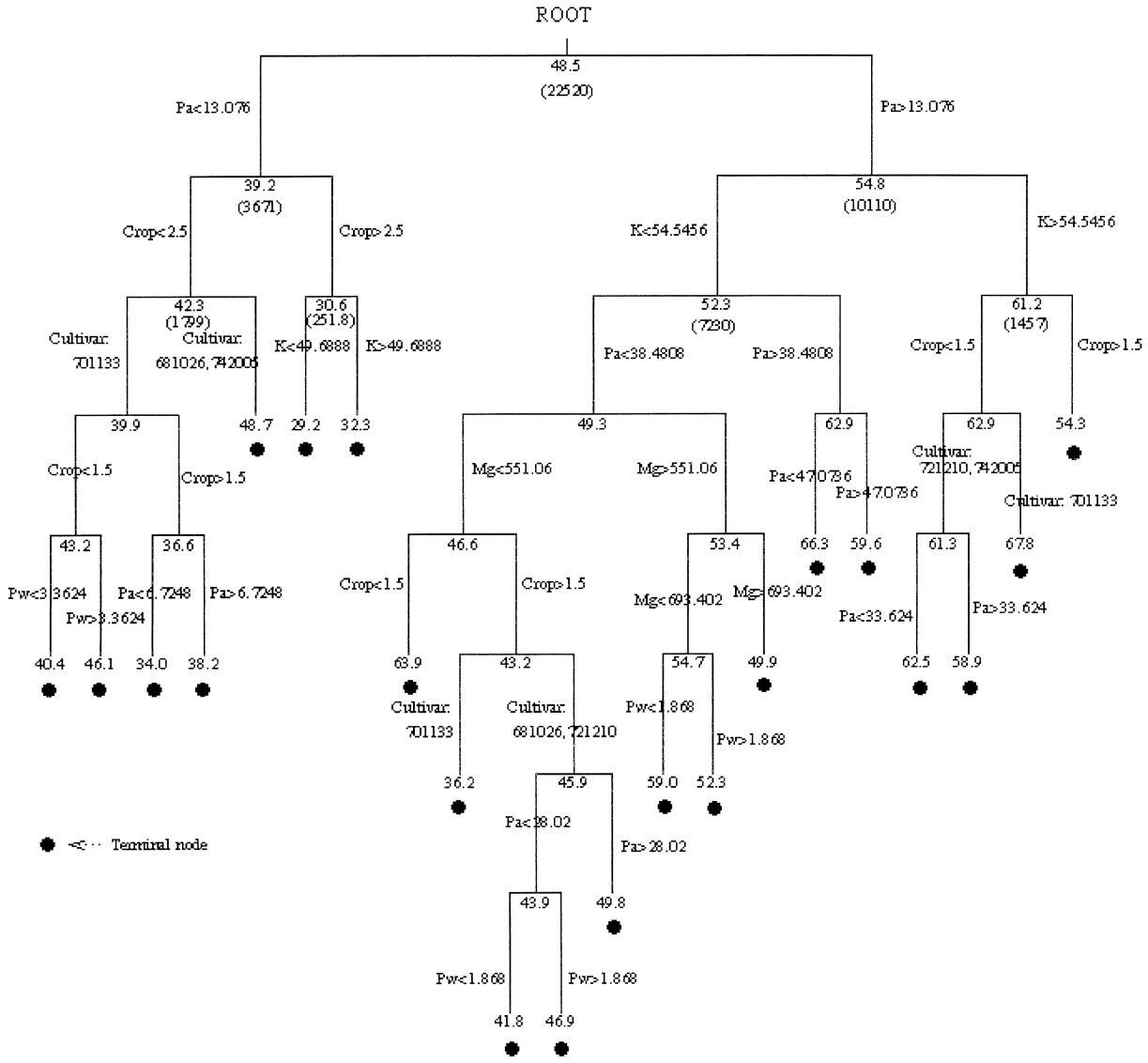

Figure 3. Maximum regression tree.

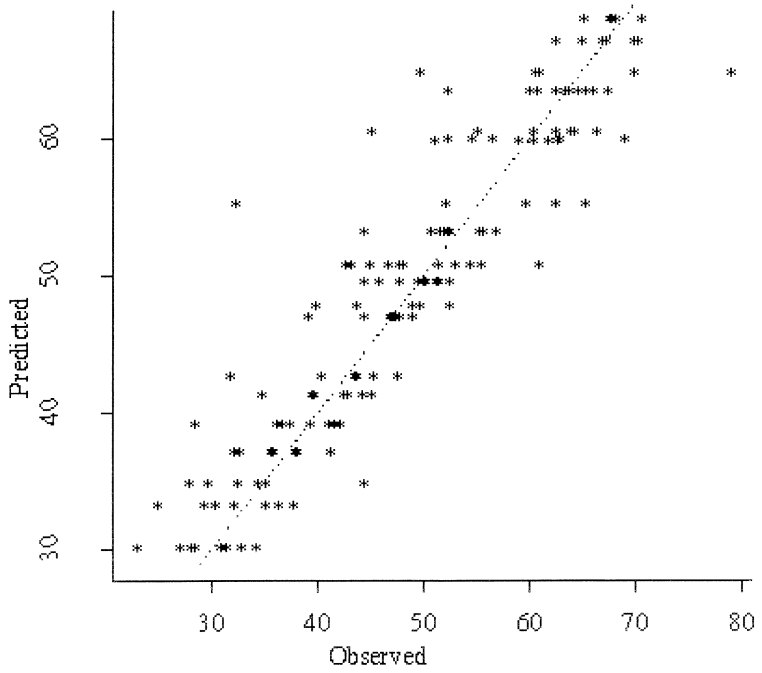

Figure 4. Observed versus predicted **TPA** yields for maximum regression tree.

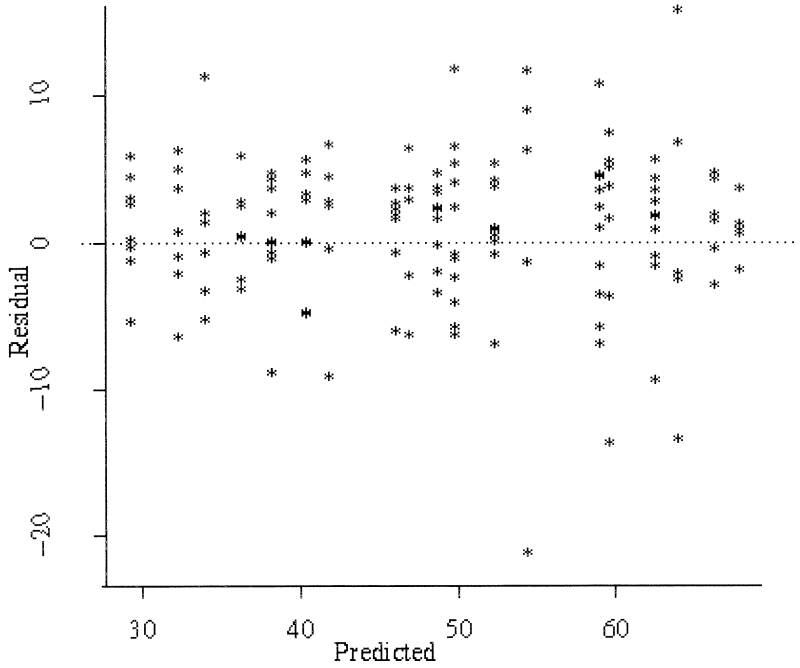

Figure 5. Residuals versus predicted **TPA** yields for maximum regression tree.

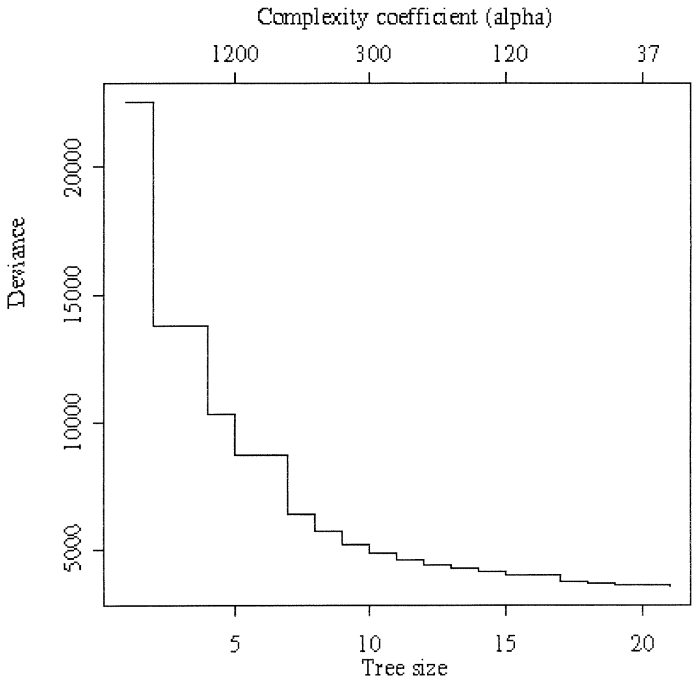

Figure 6. Pruned tree deviance versus tree size and complexity coefficient.

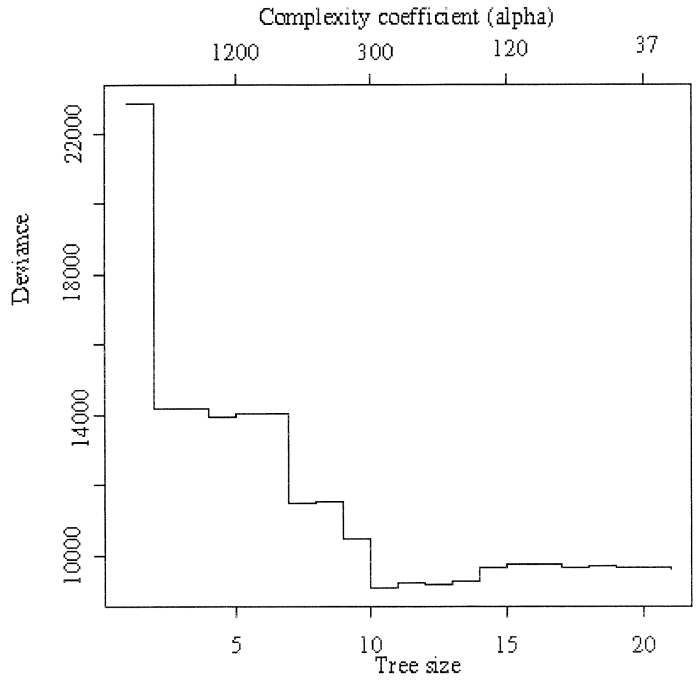

Figure 7. Deviance versus tree size and complexity using ten-fold cross-validation.

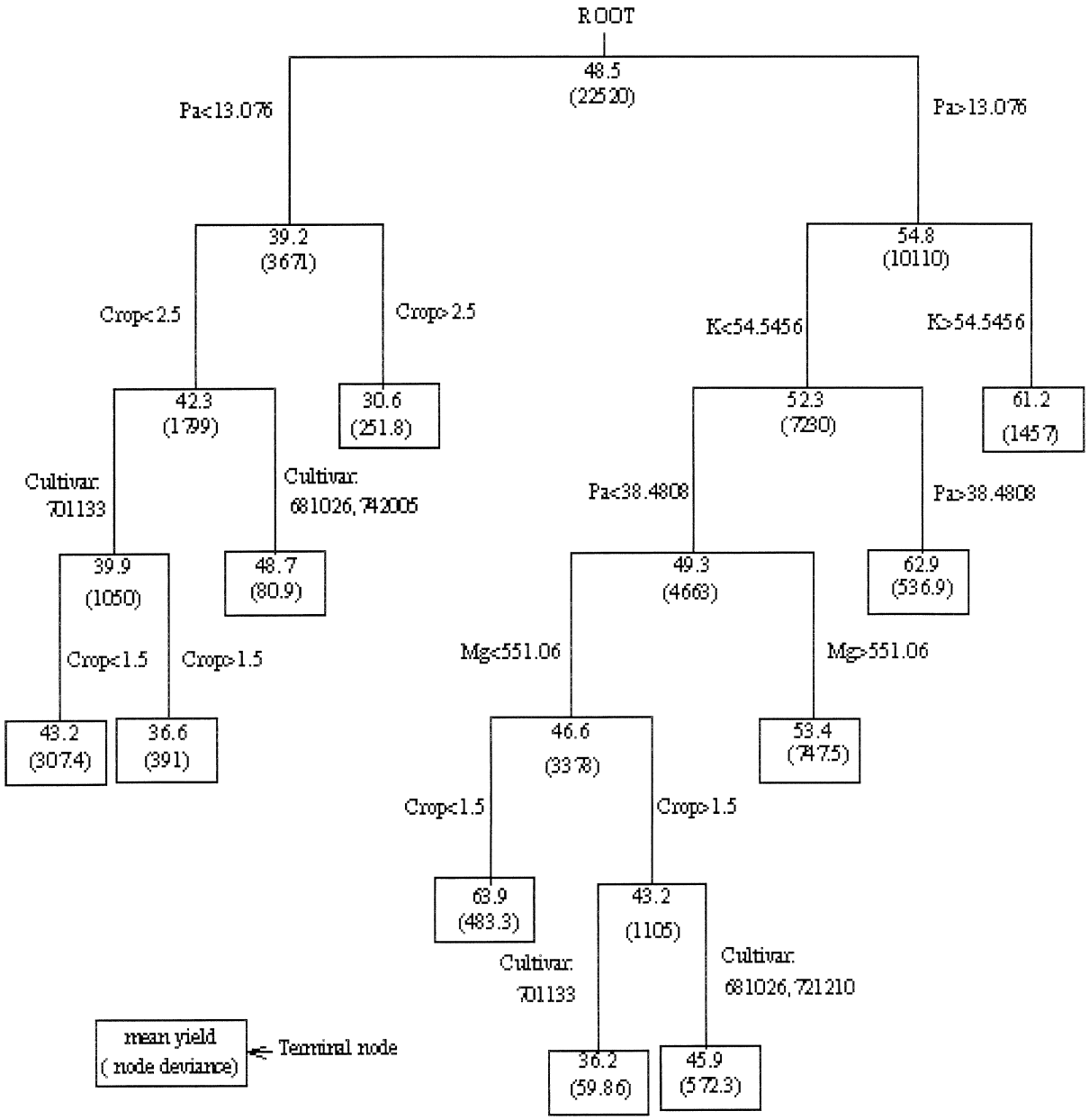

Figure 8. Optimal ten node regression tree.

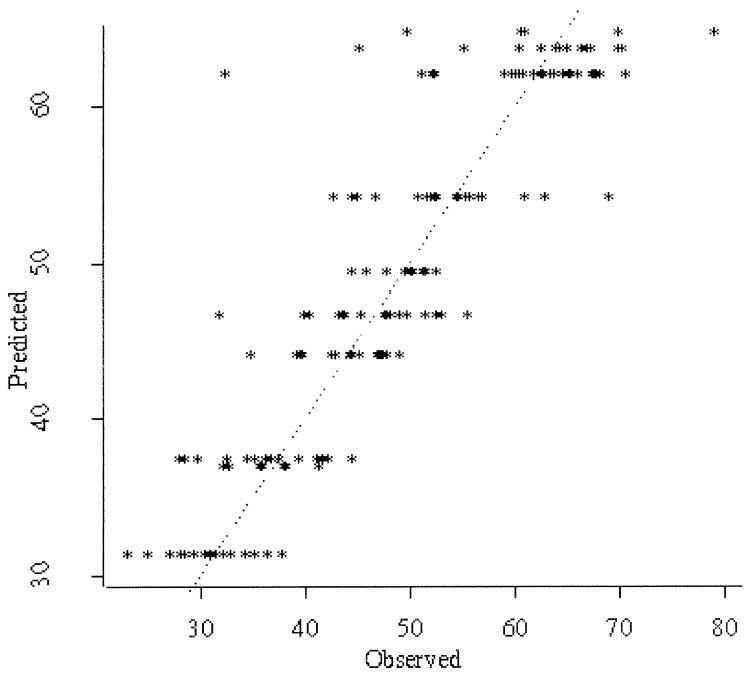

Figure 9. Observed versus predicted TPA yield for optimal ten node tree.

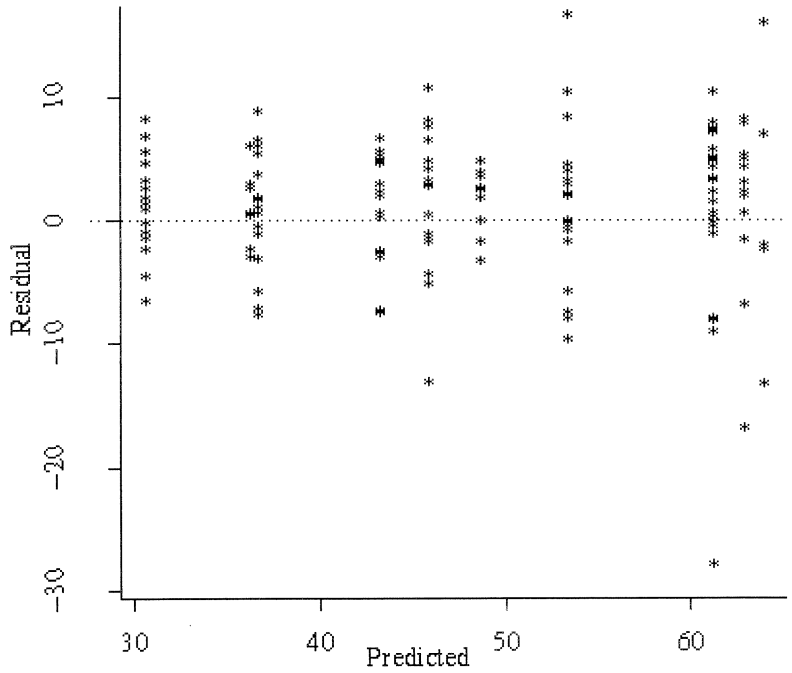

Figure10. Residuals versus predicted for optimal ten-node tree.