CARTscans: A Tool for Visualizing Complex Models

Martha NASON, Scott EMERSON, and Michael LEBLANC

We present CARTscans, a graphical tool that displays predicted values across a four-dimensional subspace. We show how these plots are useful for understanding the structure and relationships between variables in a wide variety of models, including (but not limited to) regression trees, ensembles of trees, and linear regressions with varying degrees of interactions. In addition, the common visualization framework allows diverse complex models to be visually compared in a way that illuminates the similarities and differences in the underlying methods, facilitates the choice of a particular model structure, and provides a useful check for implausible predictions of future observations in regions with little or no data.

Key Words: Bagging; Boosting; Classification and regression trees; Color coding; Graphics; Linear regression; Random forests; Visualization.

1. INTRODUCTION

This article describes graphical techniques inspired by the use of CT Scans in medical imaging. In CT scans, a series of two-dimensional images, taken in successive slices, each one right above or below the one before, is used to depict the three-dimensional structure being scanned, such as a head or body. The images are then printed in a line or a grid, and the viewer must mentally restack the images on top of each other to form a mental representation of the three dimensions. With practice, physicians can get quite adept at visualizing the size, shape, and density of a wide range of three-dimensional structures.

The tools we present take a similar approach to medical CT scans by presenting a series of "slices" of the predictor space. These tools can help visualize any space of between three and six dimensions in many contexts, linear and nonlinear. Our focus is on how these tools can aid in understanding the structure of diverse models, and facilitate the comparison of different types of models, in a way that parallels what can be seen in a simple linear regression context. We begin by showing how these tools can be used to

Martha Nason is TKKK, Biostatistics Research Branch, NIAID, 6700B Rockledge Drive, MSC 7609, Bethesda, MD 20892 (E-mail: mnason@niaid.nih.gov). Scott Emerson is TKKK, and Michael LeBlanc is TKKK, Department of Biostatistics, Box 357232, University of Washington, Seattle, WA 98195.

Journal of Computational and Graphical Statistics, Volume 13, Number 4, Pages 1–19
DOI: 10.1198/106186004X11417

visualize a relatively simple single regression tree, as popularized in the statistical literature by Breiman, Friedman, Olshen, and Stone (1984). Regression and classification trees are flexible, partition-based models that are well-suited to describing complex interactions. At their most basic, these models recursively partition the predictor space into disjoint rectangles that are successively more homogeneous with respect to an outcome variable. As the trees get more complicated, however, so does the description and interpretation; thus, we broaden our scope to demonstrate the use of these tools in understanding, visualizing, and comparing a range of models, including linear regression models with various degrees of interactions, and models comprised of aggregates of trees, such as bagged (bootstrapped aggregate, Breiman 1996) or boosted (Friedman 2001). The tools discussed here can be applied to a wide variety of models because they work only on the predictor space and the predicted values. We discuss several issues that arise in the specification of these plots, such as measures of variable importance and their function in choosing which variables to display and in what roles, and how these tools can help in assessing the variability behind the predictions in these models.

1.1 CURRENT GRAPHICAL TOOLS

There are many tools for visualizing data, some of which were described by Tukey and Tukey (1981). Although many of the concepts and ideas behind visualizations have not changed since the publication of that article, the rapid growth of computing power has led to a corresponding growth of new visualization methods. Some of these are general, like XGobi's well-known "Grand Tour" (Swayne, Cook, and Buja 1998), and some are specific to certain types of models. For tree-based models, for instance, there exist a range of tools, both commercial and free. Many of these focus on the hierarchical structure, ranging from traditional flowchart-like plots, to interactive "drill-down" tools such as those popular in commercial datamining software. Other efforts to display hierarchical structure include "Treemaps" (http://www.cs.umd.edu/hcil/treemap/), which recursively partition a viewing screen into color-coded rectangles based on the relative importance of a node. These displays, which have proven useful in visualizing file directory structure and financial databases, depend on an explicit hierarchy in the organization of the data. They have been used for statistical trees, and are useful for showing the organization and size of the leaves, but none of these plots are geared towards representing main effects and overall effects of certain variables. Furthermore, none of these plots truly achieve our goal of tools that allow an understanding of the structure of a tree, or of an ensemble of trees, that parallels the understanding of the equivalent linear model. Few visualizations give equal insight to a single regression tree, an ensemble of trees, or a linear model, in terms of the structure of the model and the corresponding insights about relationships between variables. Indeed, visualizing linear regression models has not been a main goal of visualization packages, as these parametric models are often considered well summarized by parameter estimates and standard errors.

Enhanced scatterplots (Urbanek 2002) work well for showing the relationships between

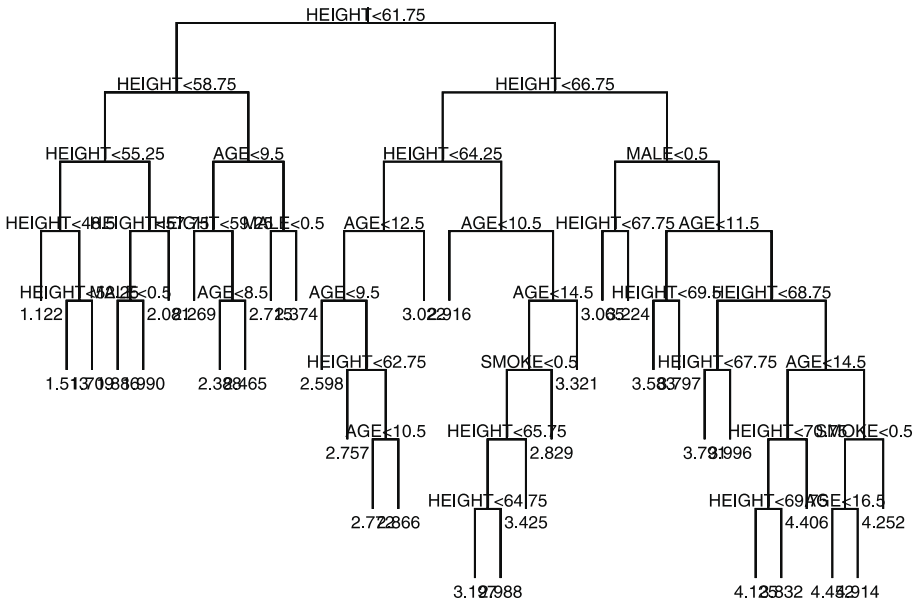

Figure 1. Hierarchical view of tree.

two variables when the model is built on only two dimensions; our methods generalize these views to four or five dimensions. There have been several attempts to allow two-dimensional visualization of functions of more than two variables, either by conditioning on all but two of the variables, or by averaging over them. Conditioning on sets of values of all but one or two of the variables leads to a set of plots known as a *trellis* (Becker, Cleveland, and Shyu 1996). Averaging across the effects of all but one or two of the predictor variables leads to a plot known as a *partial dependence plot* (Friedman 2001; Hastie, Tibshirani, and Friedman 2001). These are among the few tools that have proven useful for visualizing both single tree-based models and ensembles of trees, and could be useful for linear models. The tools presented here in many ways combine the features of the trellis plots, the partial dependence plots, and the enhanced scatterplots. We combine and extend those methods to allow increased flexibility in specifying weighting functions to represent predictor distributions, calibrating to assess variability, and visually assessing effects of the most important variables.

2. AN EXAMPLE: THE FEV DATASET

We introduce our method with a simple example from an introductory biostatistics textbook (Rosner 1990). This dataset is comprised of the age, height, sex, smoking status, and lung function as measured by forced expiratory volume (FEV) in 654 children aged 3 to 19. Our main question of interest is in the effect of smoking on FEV, allowing for and expecting correlations and interactions between the covariates. Figure 1 shows a traditional, hierarchical view of a cross-validated regression tree for this data. Although there is much

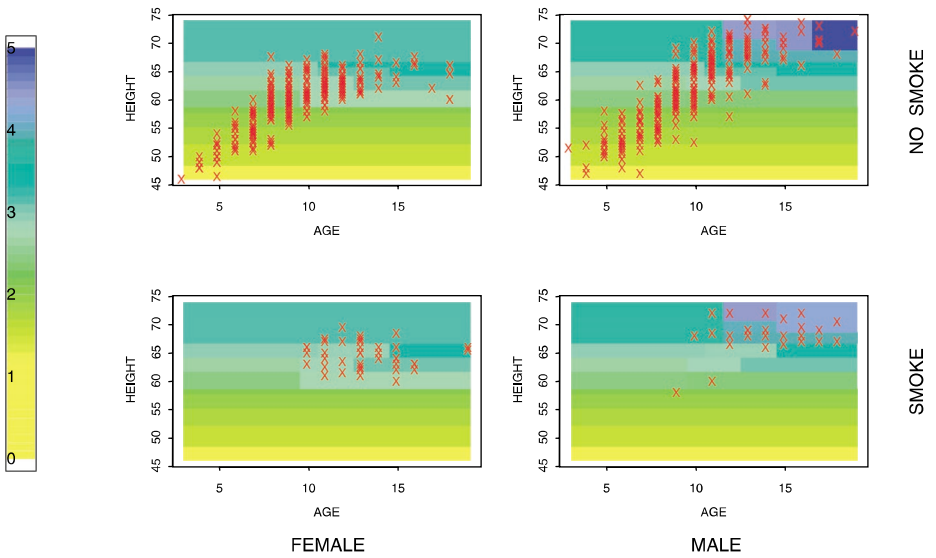

Figure 2. CARTscan of FEV tree.

information there about the structure of the model, there is much that this view does not capture. While improvements can and have been made upon this view, such as making the height of the vertical bars proportional to the change in deviance at each split, or color coding the leaves (see Figure 8, p. 16), understanding the structure of the tree, the relationships between variables and the magnitude of effects from these plots remains difficult at best. Addressing the main question of interest, when the question is the role a particular variable plays, is limited to descriptive observations about where the variable appears in the tree, how often, and with what change in deviance.

Figure 2 shows a CARTscan of the cross-validated tree on the FEV data. In this case, the CARTscan is a 2×2 matrix of stratified plots, with strata defined by the two dichotomous predictors in the data. For each combination of sex and smoking status, the predicted values are color coded and displayed as a function of age (x-axis) and height (y-axis), with the data points superimposed in red. The yellow regions represent the lowest predicted values, and the blue the highest: a range of greens codes for the values in the middle, as can be seen from the thermometer-style legend on the right side. Once accustomed to looking at these plots, several things become immediately apparent about the structure of our predictions: height is clearly the most influential predictor of lung capacity, regardless of age, sex, or smoking history, based on the horizontal bands of color ranging from yellow to blue in each plot. We see instantly that the lowest predicted FEV values occur in the youngest, shortest children, as these are the yellow regions of the plot. As these lower-left regions look the same on all four plots, we see that our model does not show an effect of either gender or smoking status on these children. In the upper right-hand corners of the four plots, however, we see that the highest predicted FEV coincides with the older, taller boys, and we see some evidence that

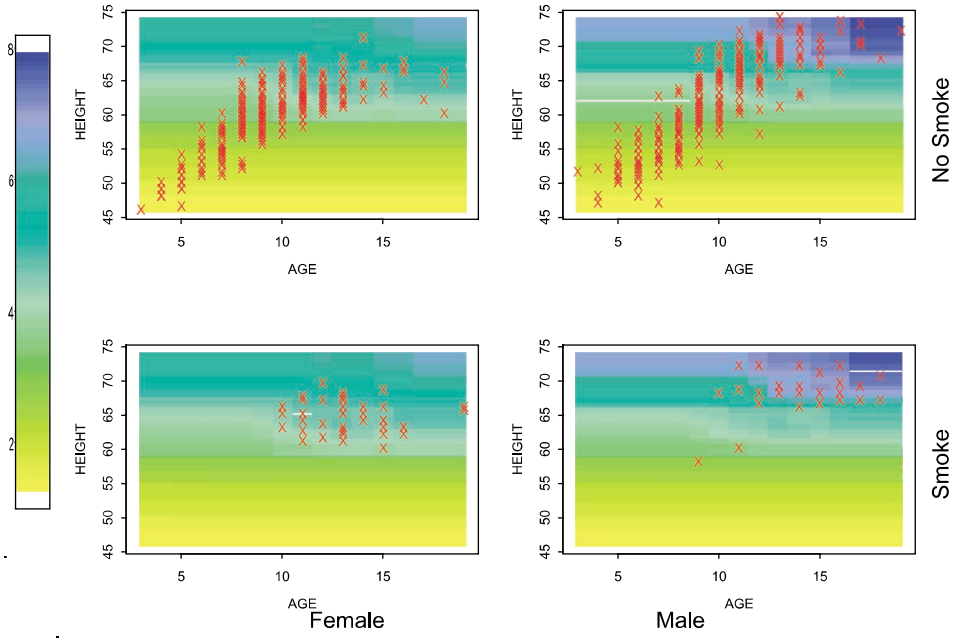

Figure 3. Bagged tree on FEV data.

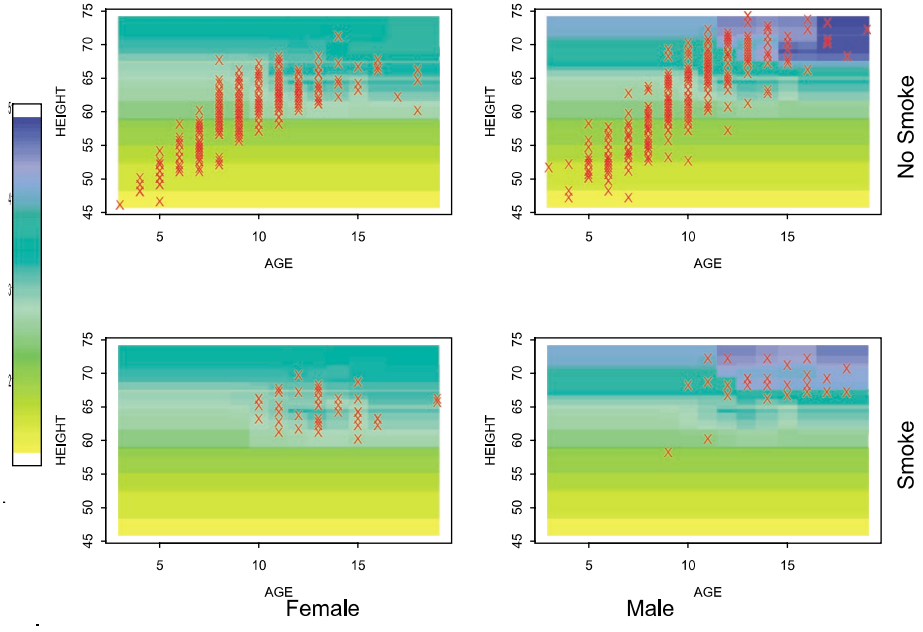

Figure 4. Boosted tree on FEV data.

there is an effect of smoking among these boys. We see that the patterns of predicted values look identical for the two sexes until about age 11, at which point the boys predicted values rise faster than the girls. None of these observations seem surprising given even the most elementary knowledge about children's growth patterns, but it is worth pointing out that we can discuss the possibility of four-way interactions here (an effect of smoking only among the taller, older boys—which, as younger boys are not as likely to smoke, may be the only age range in which we have any power to see a difference). Note that such an interaction would not show up in many linear regression models unless explicitly specified beforehand.

Regression trees are not the only predictive models in which visualization of the model structure can become difficult, or where CARTscans can be useful. Figures 3 and 4 show CARTscans of two more complicated models on these same data. Figure 3 shows a CARTscan of a bagged tree (Breiman 1996); it looks much like the original, with a few telling differences. First, the transitions between the colors have been smoothed out, blended—this is expected, as the bagging process smooths out the step functions by aggregating across different choices for the cutpoints. Second, while the general structure of the predicted values is almost the same, there are a few regions where they differ notably, such as in the older, taller girls. This is because, in some of the bootstrapped trees, there was not a split on sex, so the predictions more closely match those of the boys. This can easily happen in a region with little or no data, where the predictions are somewhat arbitrary.

Figure 4 shows a CARTscan of a boosted tree (Friedman 2001). Like the bagged tree, this ensemble method leads to a smoother-looking model. The predicted values for the older, taller girls match the original tree more closely than the bagged tree. Comparing these CARTscans with the previous one allows a quick visual comparison of the models produced by the three methods, and where their predictions are similar and different. This is a comparison which would be difficult to make otherwise, and it may be an important one in trying to decide which model to trust or in trying to understand the practical differences in the underlying methods.

Figure 5 shows two linear regressions of this data. The left panel shows all two-way interactions between predictor variables included in the model. Although this model may be considered straightforward to interpret without the CARTscans, since there are only four predictors and two are dichotomous, this view may provide extra understanding of the structure of the model for some users. The right panel takes this one step further by depicting the linear model which now includes all two-, three-, and four-way interactions between the predictor variables. For many people, this is now a difficult model to interpret when presented in a traditional way, but is included here for equivalence to the tree-based model. The CARTscan gives an elegant synopsis of the structure of the modeled relationships, allowing the user to quickly and easily evaluate the regions that where predictions are particularly high, or changing quickly, or constant over a wide range. In addition, this view allows the user to more easily assess the weaknesses of the model. For instance, in young, short girls who smoke there is a region of implausibly high predicted values from the larger model. From the difference in the range on the two thermometers at the sides we can see that these predictions are even higher than the previous model predicted for anyone. Looking

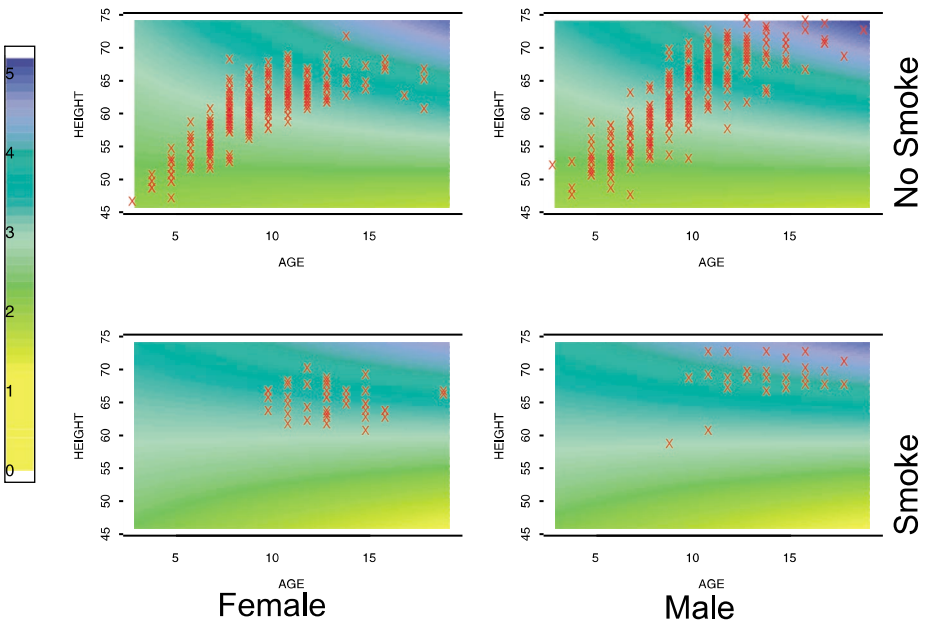

Figure 5. Linear regression on FEV data, two-way interactions.

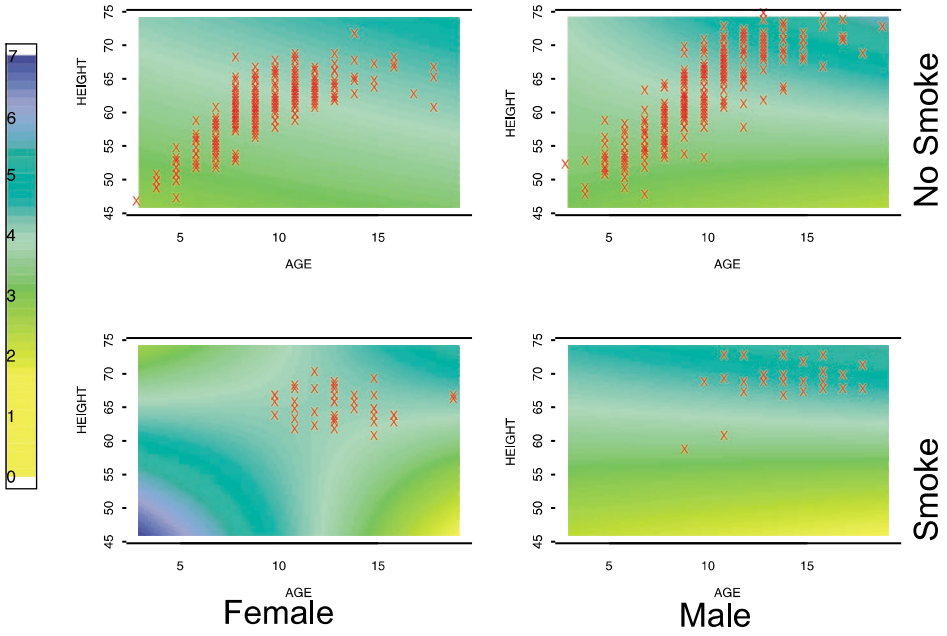

Figure 6. Linear regression on FEV data, four-way interactions.

at the plots makes it clear that this is a product of the lack of data in this region. Without the plot, however, one might be more likely to inappropriately trust a prediction of a future observation in this poorly modeled region.

With this first example as motivation, we now back up and discuss our methods in a more general framework.

3. CARTSCANS

A CARTscan is a series of graphs, each of which displays a summary measure of the response against two predictor variables, which we refer to as the "inner" variables (age and height in the FEV example). Each graph can be thought of as a two-dimensional "slice" of the higher-dimensional space in which the tree-model truly exists, based on range of values for two or three other predictors (the "outer" variables; sex and smoking status in the FEV example). By noticing the changes from one slice to the next, the user can mentally reassemble the slices into an understanding of the higher-dimensional structures. We will often refer to the predicted mean values as the response summary in the following discussion, as this is the most common measure used. However, it is worth keeping in mind that these methods will work with any summary measure, such as odds, probability, hazard ratio, or median.

As with most statistical tools, there are two categories of goals of these methods: description and inference. In both categories, there are several issues that need to be considered when using these tools, issues that can customize these methods to see different types of structures. These issues, in the order which they must be dealt with, include variable selection, choices of cutpoints, partial voluming and smoothing, and scaling. We begin with a brief discussion of variable importance.

3.1 VARIABLE IMPORTANCE AND SELECTION

In any model that contains three or more predictors, a choice must be made as to which four variables to display, and which two to display on the inner plots. Certainly this may be straightforward, as in the FEV example where there are four predictor variables and two are dichotomous, or in the case where there is one or two predictors of primary importance, in which case placing these on the inner axes will maximize the amount of information about the effect of these variables.

In most situations, however, there will be a decision to be made. If the model contains three predictor variables, only one is used for stratification; with four variables, a matrix of graphs is created where one of the outer variables varies horizontally along the page and the other changes vertically. With five variables, one option is to continue the stratification, using different pages for different levels of the fifth variable, or to condition on a region of this fifth dimension. Another option is to smooth over the effect of the fifth and higher variables, using any of a number of smoothing techniques. The exception is that if all the predictor variables are binary, it is reasonable to stratify on up to four variables per page, with four more variables used in the inner plots, allowing up to eight binary predictor variables

to be displayed on one page. If multiple pages are desired, two additional variables can be specified for this level of stratification.

There are a wide variety of approaches that can be taken in assigning the roles, ranging from choosing variables (and cutpoints for these variables) that maximize a measure of structure on the inner plots to choosing the first two variables to be split on in the tree-based model. A useful idea is to define a measure of variable importance that can capture the relevant influence of a variable on the models in question. In a simple tree model, an initial, straightforward measure of the importance of variable X_j might be formed by summing an appropriate measure of impurity $\hat{\imath}^2$ (such as the sum of squared residuals) over each split for which X_j is the splitting variable:

$$\mathrm{VI}_j^2(T) = \sum_{t=1}^{K-1} \hat{\imath}_t^2 I(v(t) = j),$$

where $\hat{\imath}_t^2$ is an appropriate measure of improvement at that node (such as change in squared error), T is a regression tree with $K-1$ interior nodes, $v(t)$ denotes the index of the variable that maximized the change in impurity $\hat{\imath}_t^2$ at node t (and thus was chosen as the splitting variable), and I is an indicator function, taking value 1 when variable j was the splitting variable at this node, and 0 otherwise. A similar approach to measuring variable importance was taken by Breiman in his article on random forests (Breiman 2001): for each variable, randomly permute the values of that variable across individuals, while holding all other predictors constant. Repredict for the data that has been "noised-up" for variable X_j, and use the change in impurity $I(X*)$ as the measure of variable X_j's importance. This conceptually simple idea has the advantage that it can be effortlessly generalized to any predictive model or aggregate of models, by averaging across the models (Hastie, Tibshirani, and Friedman 2001). Furthermore, it allows for comparing the measures of importance given by several distinct classes of models (for instance, a linear regression and a tree-based model). Note also that this method generalizes easily to permuting over sets of variables, which allows us to simultaneously explore pairs (or larger sets) of variables without losing the covariance structure in these subsets. Alternative measures, such as those based on the idea from the CART monograph of *surrogate splits*, might be useful in situations where one variable might "mask" the effect of another variable.

Our approach, as a default for CARTscans, is based on the measure of variable importance from Breiman's work on random forests. Our steps are as follows:

1. Compute variable importance measure (VI) for each predictor variable.
2. Order the variables $X_1^*, \ldots, X_p^*$ such that $\mathrm{VI}(X_1^*) \geq \mathrm{VI}(X_2^*) \geq \cdots \geq \mathrm{VI}(X_p^*)$.
3. If X_1^* has a small number of unique values (default: ≤ 5), set X_1^* as outer variable; else set X_1^* as inner variable.
4. If X_2^* has a small number of unique values (default: ≤ 5), set X_2^* as outer variable; else set X_2^* as inner variable.
5. Assign X_3^* and X_4^* to the positions not filled by X_1^* and X_2^*, such that there are exactly two outer variables and two inner variables.

We use this algorithm on the assumption that the effects of the most important con-

tinuous variables belong on the inside, where they can be seen in as continuous manner as possible. We note that this is not the only choice: it has been suggested that some people might have an easier time assessing effects if they are displayed over a larger space, and therefore the most important variables should be displayed on the outer axes; we leave this as a choice for the user.

We note that we often find it useful to iterate through combinations of variables, rotating which appear as inner variables, which as outer, and which are smoothed across. This iteration can often be a very useful tool regardless of variable importance measures, and can show unexpected structures from angles that might have been missed.

3.2 CUTPOINTS

If the outer variables are dichotomous, or have a small number of possible values, the choice of cutpoints for the outer strata is straightforward. When the outer variables are continuous, however, it is necessary to choose the number and location of cutpoints. Our default is to choose linear cuts, in the belief that it is easier to understand structure on a linear scale. An algorithm implemented in the CARTscans code attempts to minimize the amount of smoothing necessary over a small number of cuts on each outer variable. We compute, for a given number of evenly spaced splits on each outer variable, the "displayed" value $\hat{y}_i^{\text{displayed}}$ for $\vec{x_i}$ where $\vec{x_i}$ are either on a grid over the predictor space, or the original data points. The calculation of the displayed value $\hat{y}_i^{\text{displayed}}$ will depend on the choices of weighting and smoothing functions, as discussed in the following section.

Our algorithm then minimizes the sum of squares $\sum_i (\hat{y}_i - \hat{y}_i^{\text{displayed}})^2$ over a range of numbers of cutpoints for each of the two variables, and chooses the smallest number of strata within the range that minimizes this statistic.

There are plenty of other options for choosing cutpoints, and we leave these as choices for the analyst. Cutpoints can be based on the quantiles of the predictors, or on the first thresholds used in the tree model. Other options, whose viability may depend on the size of the dataset and the available computing power, include optimizing some summary measure of the structure by looking at all possible cutpoints, or building a new tree with only the outer variables, either one or two at a time, and a large minimum node size to choose a small handful of thresholds.

It is worth noting that if there are only four or five predictors, it is possible to avoid smoothing altogether by using all thresholds for all splits anywhere in the tree on the outer variables as the strata-defining cutpoints. In most situations, however, this will lead to a very large number of plots, making it difficult to decipher the structure.

Another option is to have the strata overlap. This may be useful when the cutpoints are not chosen for scientific reasons.

3.3 PARTIAL VOLUMING AND SMOOTHING

If the outer variables are continuous, it is likely that there are different values of these outer variables which lead to different predicted values but which will be displayed together

in an inner plot. In a model such as regression or a set of bagged regression trees this is inevitable; in a single regression tree, it remains likely that at least one split on an outer variable is not represented by an outer cutpoint. Therefore some or all regions of some plots will have more than one predicted value to represent. This is also inevitable if there are more than four or five variables split on in our tree-based model.

This is referred to as "partial voluming," a term from the CT scan culture, referring to the appearance of a structure in varying degrees on several successive slides. When this happens, we might see some "ghosting," where a faint color change can indicate either many points with a slightly different predicted value or a few points with a drastically different predicted value. This is one of the times that iterating through the plots to put different variables in the inner, outer, and smoothed positions can be important.

It is recognizing the issue of partial voluming that a smoothing or weighting function needs to be specified. Different choices can lead to different pictures, especially when the displayed predictors are correlated with one or more undisplayed but relevant variables. One obvious choice is to use weighted averaging with the weighting based on the Euclidean size of the regions. We refer to this as uniform weighting because it is equivalent to assuming a uniform distribution of all predictor variables. Another choice is to weight based on the empirical distribution of the predictors, with or without kernel estimation. Other choices are possible, such as specifying a uniform distribution for certain predictors and an empirical one for others, or using a kernel density smooth. These choices allow flexibility in generalizing to different predictor distributions. This is important, because the trees present "averages" over potentially large regions. If the underlying model is only approximated by the tree (say some smooth function) then the predicted average may deviate from the average for a new sample in the same region (node) if the underlying covariate distribution changes.

For a tree-based model, a reasonable set of partitions can be chosen for the inner variables based on the splits in the tree or trees. For other types of models that do not have inherent partitions, a grid over the range of the inner variable is a useful choice. Clearly, the finer the grid, the more closely the final set of images can represent a smooth underlying function. The displayed value for each inner region must be averaged over all possible predicted values for the appropriate range of outer variables, and for any undisplayed variables. For uniform weighting, this average is weighted by the size of the regions being averaged. For empirical weighting, this is a weighted average of the predicted values of the observations that fall in this region. For example, the left plot in Figure 7 shows a hierarchical view of a regression tree of $Y \sim X1 + X2 + X3$, where $X1$ and $X3$ are correlated, and all three predictor variables are split on in the tree model. For $X2 \leq .5$, there is only one split on $X1$, at approximately .49. When uniform weighting is used in the first CARTscan, there is no sign of the grid used to predict or the influence of $X3$, but these start to show if empirical weighting is used, on the right-hand CARTscan. The CARTscan built with uniform weighting displays only two regions for $X2 \leq .5$, whereas the plot where empirical weighting is used, appears to show more regions in this bottom half of the CARTscan: this is because the points with lower values of $X1$ also tend to have lower values of the undisplayed variable $X3$, and as a result the average predicted value of the points in the region $R = \{x1, x2 : x1 \leq .32 \text{ and } x2 \leq .5\}$ is smaller than the average

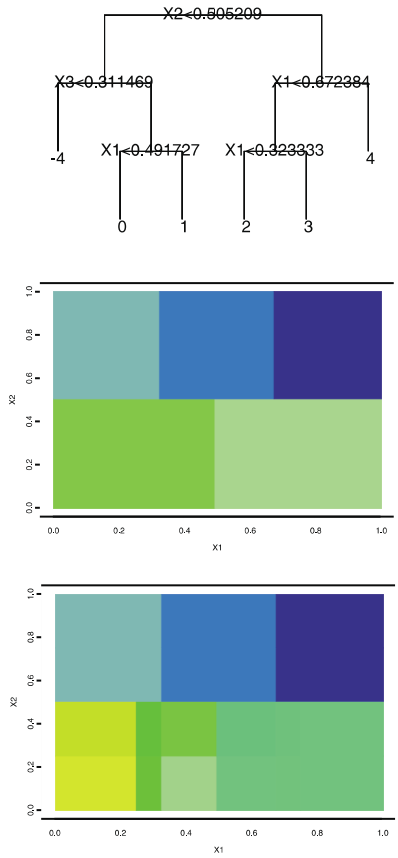

Figure 7. Hierarchical view and CARTscan of ghosting example, with important undisplayed predictor variable: uniform (middle), and empirical (bottom) weighting.

of the points in the region $R = \{x1, x2 : .32 \leq x1 \leq .49 \text{ and } x2 \leq .5\}$. This "ghosting" effect, where the correlations cause regions to appear that are truly due to the effect of undisplayed values, can be considered undesirable, as it may produce misleading ideas of how the model splits on the displayed variables, but it can also be considered a resource, as differences between the uniform and empirically weighted CARTscans can serve as a flag that there are important undisplayed predictor variables and correlations. In addition, the two options allow something of a continuum between displaying a model independent of the data that generated it (uniform), and displaying a data space as represented by a model (empirical). In the latter case, the ghosting may be totally appropriate as it does indicate differing predicted values in the region.

3.4 SCALING AND COLOR CODING

The decisions about how to color code the regions are in essence questions of how to calibrate the response distribution, and are equivalent to the choices on the range of an axis

in a scatterplot: changing the range on the vertical axis can for instance make a line appear practically flat or steeply sloped. The choices include setting sample-based, scientifically relevant absolute limits, setting descriptive limits such as the range of the predicted values, or setting limits based on point-wise confidence bands, a choice which brings an element of inference into the displays.

One choice, and a reasonable default, is to simply set as yellow the lowest predicted value from the model and blue as the highest. In the case of two predictor variables, this will assure at least one yellow rectangle and at least one blue one.

If the highest or lowest predictions occur only for data with a small range of values for a third variable which has been smoothed over, the smoothing or averaging may disguise the extreme colors and lead to a graph with only greens appearing. A similar phenomenon may occur if the third variable has been used for stratification instead of smoothing and the extreme values occur over a range of this variable smaller than the binning imposed by the strata. For more discussion of the smoothing, see Section 3.3.

An alternative is to set the color range based on only *displayed* predicted values. This option guarantees use of the full range of colors for the matrix of plots, but may mislead the user into thinking that the extreme values are in a certain region when in fact the extreme values depend on variables or regions averaged across.

As always, if there is a scientific reason that only a range of the predicted values is of interest, the range of predicted values to be plotted may be specified. Summary measures that fall outside this range may be left the color of the background, or all values below the range may be set to yellow and all values above the range to blue. This technique, called "windowing" in the medical CT scan culture, allows for more flexibility in focusing in on a specific region of predictions. For instance, it is often the case that the interest lies in identifying regions with very high predicted values, and it will therefore be useful to set the lower end of the range to be a high value, so that all cases with a low to medium predicted value show up as yellow, and green and blue concentrate on the structure of the regions with high predicted values.

Another option is to calibrate the colors to the range of predictions to point-wise confidence intervals, based on a set of bootstrapped trees. This can be a very useful tool, because if there is wide variation in the predictions the original numbers will all be coded as middle greens, as the bootstraps are bound to have lower minima and higher maxima than the original model. This calibration can function as a sort of standardizing, and can facilitate a sense of what apparent structures "wash out" in bootstrapping. Further discussion of the bootstrapping and recalibration can be found in Section 3.6.

We also wish to point out that our choices for the range of colors are far from unique. We found these colors to be useful to a variety of users—including one who was color blind—but others may wish to make other choices. Brewer (1999) included a discussion of color perception and choice, as well as a useful piece of software designed to aid in these choices. In addition, there has been research into linear color scales—that is, choices for color spaces that are perceived linearly. We note that the color scheme used here is close to linear, but is not exact.

3.5 DISPLAY

Once the outer and inner variables have been chosen, as well as the cutpoints for the outer variables, a matrix of plots is produced. Without loss of generality, let X_1 and X_2 be the inner variables, X_3 and X_4 be the outer variables, and $X_5, \ldots, X_p$ be the variables that are smoothed over. Denote the cutpoints for X_3 as $C_{3,1}, C_{3,2}, \ldots C_{3,m_3}$, and for X_4 as $C_{4,1}, C_{4,2}, \ldots C_{4,m_4}$. Then the plots are stratified with the first plot showing the predicted values (or other summary measure) as a function of X_1 and X_2 for values of $X_3 \epsilon \{C_{3,1}, C_{3,2}\}$ and $X_4 \epsilon \{C_{4,1}, C_{4,2}\}$. Continuing along the top row, the range would not change for X_3, but would step up through the intervals for X_4, starting with $X_4 \epsilon \{C_{4,2}, C_{4,3}\}$. Similarly, the rows would contain successively higher intervals for X_4, ending with the final plot of $X_3 \epsilon \{C_{3,m_3-1}, C_{3,m_3}\}$ and $X_4 \epsilon \{C_{4,m_4-1}, C_{4,m_4}\}$. Within each stratum, the inner variables and the desired summary measure, such as the predicted values, are plotted against each other. Regions of the plot are shaded to indicate levels of the summary measure. Specifically, areas in which the summary measure is highest are shaded dark blue, and areas in which it is lowest are shaded yellow. All intermediate values show up as combinations of these two colors, spanning through the greens.

There are several features that can extend the graphical impact of these plots. For example, superimposing the data points that fall within each stratum on each plot as a set of red X's can add information about the predictor distribution. This is particularly useful when the data are correlated or nonuniformly distributed, as in these cases there may be boxes that appear in the plots that have no data in that box in that stratum. Thus, a sense of where the data lie can help an investigator avoid drawing conclusions about a pattern in a stratum, when that pattern is in truth a result of the data in a separate stratum, as was demonstrated in Figure 6 (p. 7).

3.6 BOOTSTRAPPING

It can be difficult with a tree-based model to get a sense of the variability of the predictions. In order to get a sense of this variability, it can be very helpful to examine point-wise bootstrapped confidence bounds. First, a bootstrapped dataset is created, based on any of the different bootstrapping algorithms found in Efron and Tibshirani (1993). For the examples in this article that include bootstrapping, we resampled the residuals across the entire dataset. Once a new dataset has been generated, a new tree is grown in the same way as the original, including any pruning or shrinking criteria. Now the weighting scheme becomes important again: a new predicted value is obtained, either for every observation (empirical weighting), or for observations on a fine grid (uniform weighting). This process is then repeated, and the predicted values are aggregated into a set that is available in the end for point-wise, or region-wise, confidence bands.

At this point, a new set of plots can be generated, one of which shows all the regions colored according to the lower quantile of their predicted values, another of which depicts the higher quantile, and a third showing mean or median values. We note that plotting the mean of the bootstrapped predictions for each region gives a view similar to that of the

corresponding bagged tree model (Breiman 1996), differing only in that these displayed regions are conditional on the original tree. Figure 9 shows an example of these bootstrapped confidence plots for a simple example.

In the low-noise situation, the predicted values will remain relatively unchanged in the majority of the bootstrapped trees, and the lower and upper pictures will look similar to the original, continuing to show the structures that were seen in the first. In the case of high variability, however, the lower plot will be primarily yellow and the upper one primarily blue; if the colors for the original are calibrated or recalibrated to the range from the bootstrapped predictions, this original tree will show as solid green. If some regions of the tree are more stable than others, these areas may hold their color in all three plots while other areas will "wash out."

We wish to emphasize that we make no claim as to the coverage properties of these confidence intervals as a whole; they should be thought of as point-wise intervals, whose purpose is to allow us a means of capturing some sense of the variability in the predictions.

4. EXAMPLES

4.1 A SIMPLE REGRESSION TREE EXAMPLE

For a demonstration, we adapt data used by Chipman, George, and McCulloch (1998) for their Bayesian trees. This model is appealing because, while its structure is simple, "it tends to elude identification by the greedy algorithm which chooses splits to minimize residual sums of squares ... The greedy algorithm is unable to capture the correct model structure" (Chipman, George, and McCulloch 1998). So while the data look perfect for a CART model, there are hidden difficulties.

For this simulated example, let X_1 be uniformly distributed between 0 and 10, and let X_2 be equal to the integers 1, 2, 3, 4 with equal probability. The data for this example come from the model $Y = f(X_1, X_2) + 2\varepsilon$ where $\varepsilon \sim N(0, 1)$ and

$$f(X_1, X_2) = \begin{cases} 8 & \text{if} \quad X_1 \in [0,5] \quad \text{and} \quad X_2 \in \{1,2\} \\ 2 & \text{if} \quad X_1 \in [5,10] \quad \text{and} \quad X_2 \in \{1,2\} \\ 1 & \text{if} \quad X_1 \in [0,3] \quad \text{and} \quad X_2 \in \{3,4\} \\ 5 & \text{if} \quad X_1 \in [3,7] \quad \text{and} \quad X_2 \in \{3,4\} \\ 8 & \text{if} \quad X_1 \in [7,10] \quad \text{and} \quad X_2 \in \{3,4\} \end{cases} .$$

These data are hard to fit with traditional trees that look marginally. The authors discussed how the greedy algorithm splits on X_1 first more often than on X_2, and consider this a weakness of the traditional greedy algorithm.

Figure 8 shows a traditional schematic of a tree grown using the greedy algorithm and pruned using ten-fold cross-validation; as a small step toward making it easier to visualize the structure, the nodes have been color coded according to our scheme where yellow represents the lowest predicted values and blue the highest. Chipman et al.'s (1998) complaint that these algorithms are likely to split first on X_1 holds true, and an understanding of the underlying

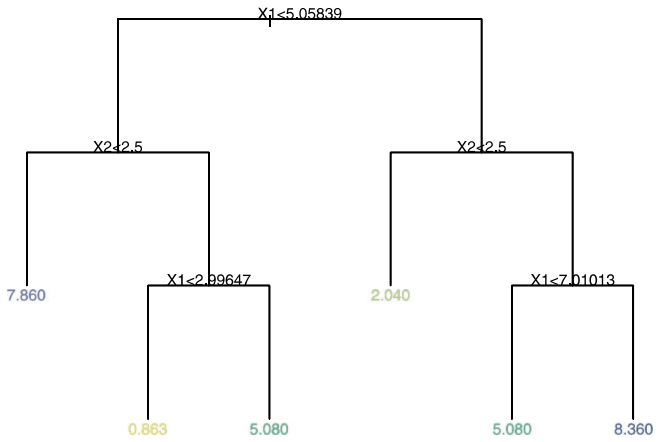

Figure 8. Tree from example from Chipman, George, and McCulloch (1998).

structure requires at least a careful scrutiny. Notice the two leaves whose predicted values are 5.08; these look physically distant from their places on the schematic, but in truth they are adjoining regions, with $X_2 > 2.5$ and either $2.99 < X_1 \leq 5.06$ or $5.06 < X1 \leq 7.01$. The fact that these are two separate leaves at all is unfortunate, and is the weakness to which Chipman et al. (1998) were referring. As is shown in the following, one of the major features of our method is that mistakes such as this in the tree structure are made largely irrelevant.

The first plot in Figure 9 shows a CARTscan of this tree, and clearly emphasizes the strength of these tools in visualizing the structure of the tree. In one glance, the structure is clear and visible. Y is high when both X_1 and X_2 are low or high; when X_2 is high and X_1 is mid-valued, Y is also mid-valued. And, even though the tree did indeed split first on X_1, the color coding has allowed two adjacent regions to be visually "recombined," as

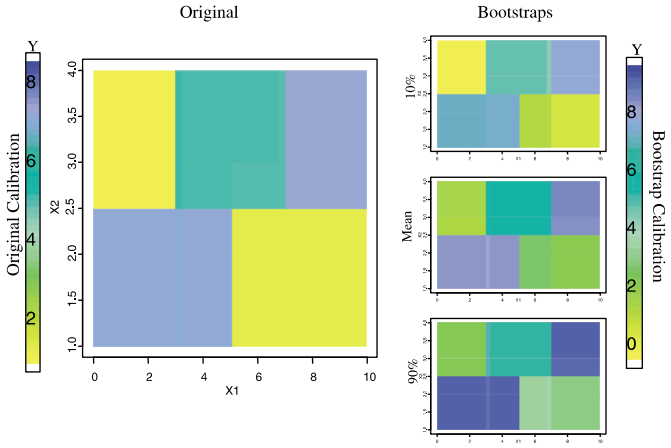

Figure 9. CARTscan of Example from Chipman, George, and McCulloch (1998).

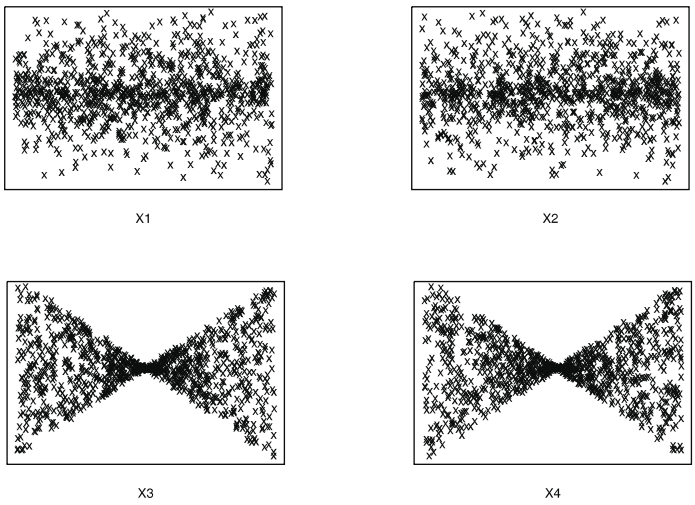

Figure 10. Residuals from linear model versus four predictors (two-way interaction).

their predicted values are indistinguishable. The green rectangle in the upper central part of the plots is much more representative of the true structure than of the modeled break at $X_1 = 5$. The graphical displays allow an understanding of the structure that the tree itself, no matter how accurate, does not depict readily; it is easy to see which areas have similar predicted values, even when the leaves are not contiguous on the trees themselves.

The second plot in Figure 9 has the colors recalibrated to the range of predicted values from 100 bootstrapped trees, and the remaining two graphs show the 10th and 90th percentiles point-wise for these bootstrapped predictions. Notice how all four graphs look basically the same; this is the graphical equivalent of a narrow confidence interval, and shows that the variability was fairly low in this example.

4.2 An Application: Residual Exploration

As in the previous example, we start with four predictor variables, independent and uniformly distributed on the interval $[0, 10]$. Let $y^* = X_1 + X_2 + (X_3 * X_4) + \varepsilon$ where

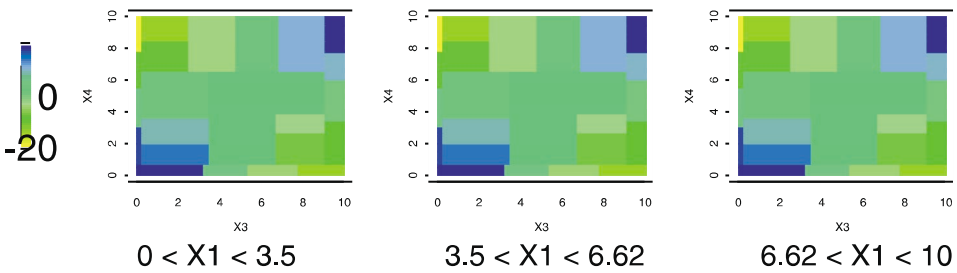

Figure 11. CARTscan of tree on residuals, two-way interaction.

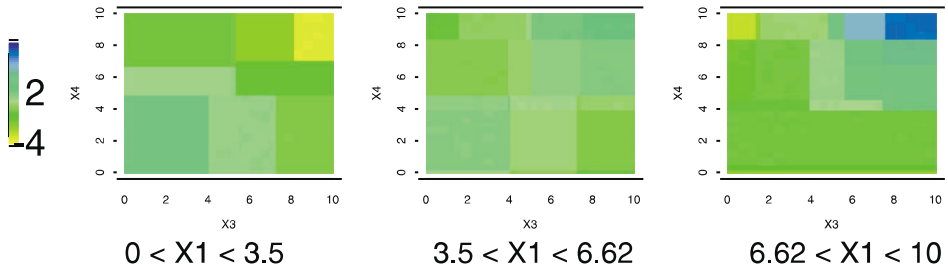

Figure 12. CARTscan of tree on residuals, three-way interaction.

$\varepsilon \sim N(0,1)$. Now suppose we have done a linear regression allowing for only main effects; let our outcome variable y be the residuals from this regression, so that $y = y^* - (\beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + \beta_4 * X_4)$ where $\vec{\beta}$ is chosen to minimize the residual sum of squares.

Figure 10 shows the residuals plotted against each of the four predictor variables in a traditional manner. Although it is clear that there is a relationship that has not been correctly modeled between X_3 and the residuals, and between X_4 and the residuals, it may be difficult to combine these pictures in to an understanding of the three-dimensional relationship between y, X_3 and X_4.

Figure 11 shows a CARTscan of a tree grown on this data, with the residuals as the outcome. Here, X_3 and X_4 are the inner variables, and only X_1 has been chosen as an outer variable; when fewer than four variables are influential in the model, it may sometimes be useful to choose one or zero outer variables. A saddle-shape becomes clear: the residuals are high when either both X_3 and X_4 are high or both X_3 and X_4 are low, they are large negative numbers when one of X_3 and X_4 is high and the other low, and they are near zero when either predictor is in the mid-range. Figure 12 shows a CARTscan of a tree similarly built on residuals, but this time when X_1 is also included in the underlying interaction. As we would expect, the residuals are largest where all three variables are either high or low.

5. CONCLUSION

CARTscans are clearly a useful set of tools for many aspects of data analysis and exploration. They allow visualization of the structure of tree-based models, linking the traditional representation of a tree to a depiction in the predictor space. CARTscans are helpful as diagnostic tools for residual exploration, and show promise for extending into measures of variable importance in tree models, a concept not easily derived from traditional graphics which focus on individual splits rather than relationships between variables. These tools are useful not only for data depiction but also for model representation. If uniform weighting is used, for instance, the images show the structure of the predictions, rather than the observations. The images can be used for prognostic rules in medicine, for instance,

regardless of how they were derived, and may allow a physician to carry a mental image of predicted risk based on four patient characteristics.

The use of these tools is in no way limited to what we have described here. These tools are very flexible, and will easily generalize to, for instance, other types of models. Much work is still to be done to flush out the full potential of these tools, and their application to new contexts will undoubtedly spur new directions for exploration.

[Received May 2002. Revised September 2003.]

REFERENCES

Becker, R., Cleveland, W., and Shyu, M. (1996), "The Visual Design and Control of Trellis Display," *Journal of Computational and Graphical Statistics*, 5, 123–155.

Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 24, 123–140.

——— (2001), "Random Forests," *Machine Learning*, 45, 5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984), *Classification and Regression Trees*, Monterey, CA: Wadsworth and Brooks.

Brewer, C. A. (1999), "Color Use Guidelines for Data Representation," in *Proceedings of the Section on Statistical Graphics*, Alexandria, VA: s American Statistical Association, pp. 55–60.

Chipman, H., George, E., and McCulloch, R. (1998), "Bayesian CART Model Search" (with discussion), *Journal of the American Statistical Association*, 93, 935–960.

Efron, B., and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, New York: Chapman and Hall.

Friedman, J. (2001), "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, 29, 1189–1232.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, New York: Springer-Verlag.

Rosner, B. (1990), *Fundamentals of Biostatistics* (3rd ed.), Belmont, CA: Wadsworth.

Swayne, D. F., Cook, D., and Buja, A. (1998), "XGobi: Interactive Dynamic Data Visualization in the X Window System," *Journal of Computational and Graphical Statistics*, 7, 113–130.

Tukey, P. A., and Tukey, J. W. (1981), "Graphical Displays of Data Sets in 3 or More Dimensions," in *Interpreting Multivariate Data*, ed. V. Barnett, Chichester, UK, pp. 189–275.

Urbanek, S. (2002), "Different Ways to see a Tree—KLIMT," in *Proceedings of the 14th Conference on Computational Statistics*, pp. 303–308.