# Extracting Representative Tree Models From a Forest.

**Hugh A. Chipman, Edward I. George and Robert E. McCulloch**

Working Paper 98-07
July 1998

Department of Statistics and Actuarial Science
University of Waterloo

ABSTRACT

A common criticism of many methods for constructing tree models is that a single tree or nested sequence of trees is produced, and that much uncertainty about the tree structure is ignored. Recent search algorithms (bumping, boosting, simulated annealing, MCMC) address this problem by finding a much richer collection of trees. They lead to an embarrassment of riches, in that it may be difficult to make sense of the resultant forest. Quite often, the problem may not be as bad as it seems: although hundreds of distinct trees are identified, many will differ only at a few nodes. Other trees may have different topologies, but produce similar partitions of the predictor space. By defining several distance metrics on trees, we summarize a forest of trees by several representative trees and associated clusters. A new plot, the *added tree plot* is introduced as a means to decide how many trees to examine while simultaneously adjusting for the goodness-of-fit of the trees considered.

Keywords: added tree plot, MCMC, tree metrics, multidimensional scaling, clustering, stochastic search, Bayes, bootstrap.

---

Hugh Chipman is Assistant Professor of Statistics, Department of Statistics and Actuarial Sciences, University of Waterloo, Waterloo, ON N2L 3G1, hachipman@uwaterloo.ca. Edward I. George is the Ed and Molly Smith Chair in Business Administration and Professor of Statistics, Department of MSIS, University of Texas, Austin, TX 78712-1175, egeorge@mail.utexas.edu, and Robert E. McCulloch is Professor of Statistics, Graduate School of Business, University of Chicago, IL 60637, rem@gsbrem.uchicago.edu. This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada, NSF grant DMS 94.04408, Texas ARP grant 003658130, and research funding from the the Graduate Schools of Business at the University of Chicago and the University of Texas at Austin.

# 1 Introduction

Recent research in tree models has produced an embarrassment of riches. By using any one of several different techniques, an analyst is able to generate a number of tree models that describe the same data set well. In some situations, these models can be quite different, while in other cases, they may be small variations on a basic structure. By defining measures of dissimilarity on tree objects and grouping together similar trees, this paper proposes a method for making sense of this "forest" of trees.

The data of Wolberg and Managasarian (1990) provides an illustration of the problem of multiple trees. The goal is to classify breast tumors as malignant or benign. Nine cellular characteristics, all measured on a 1-10 scale, are available as predictors. Additional detail is provided in section 4, and Chipman, George, and McCulloch (1998). Bootstrap (Tibshirani and Knight 1995) and Bayesian (Chipman et. al. 1998) tree search algorithms identified many promising trees; four distinct trees from this forest are displayed in Figure 1. Majority classes are indicated by "B" and "M" in terminal nodes, and misclassification rates are given below terminal nodes. All interior nodes have a split involving a single variable that results in two child nodes. Considerable differences are discernible in the four models: the first variable split upon is different in each tree, and tree 6 splits first on $X_3$ while tree 1 does not use this variable.

Over 100 other trees were also identified that fit the data well. How were these four selected as "interesting"? Until recently, common practice has been to produce plots like those in Figure 1 for a number of trees with good fit, and identify common structure by examining the pictures. A more automatic and quantitative approach is proposed here, in which trees are clustered according to several metrics.

In section 2, several methods for producing a forest of trees are reviewed. Section 3 discusses several measures of dissimilarity for trees, and in section 4 an examples is given to illustrate how these metrics may be used to cluster trees.

# 2 Methods for generating trees

In many applications, there may be many different trees that can explain the same data well. Finding such trees is a challenging problem, and one that is not satisfactorily addressed by the commonly used "greedy" algorithm (ie forward stepwise search). At each step in this algorithm, every possible split at every terminal node is considered, and the node, variable and split rule which maximize homogeneity of the two resultant children is chosen. The algorithm is only locally optimal, as splits are chosen to maximize homogeneity at the next step only.

Many improvements to this algorithm involve either manipulation of the training data or modification of the search method. Two approaches from each group are discussed below.

Breiman (1996) and Tibshirani and Knight (1995) propose random manipulation of the training data via the bootstrap (called "bagging" and "bumping" respectively). A large number of pseudo datasets are generated by resampling the original observations with replacement. When a greedy search is applied to each pseudo dataset, different trees result, some describing the original data better than a greedy tree grown to the original data. By perturbing the data, the greedy search identifies different trees, some of which may be close to a global or local maxima.

Freund and Schapire (1996) propose an algorithm (called "boosting") in which the data are iteratively reweighted instead of randomly resampled. The algorithm alternates between fitting a tree (with greedy search) and reweighting the data. The weights are adaptively chosen, with more weight given to observations that the tree models poorly. Again, a forest of trees result. Quinlan (1996) describes the application of boosting and bagging to trees.

Breiman (1996) and Freund and Schapire (1996) produce predictions for new cases by using averages of all models identified. All interpretability of the resultant model is lost, since a mixture of trees is no longer a tree. We follow more the approach of Tibshirani and Knight (1995), who use such methods to generate trees, from which one (or more) good trees are selected.

The second group of algorithms introduce a stochastic element to the search rather than manipulating the data. Chipman et. al. (1998) and Denison, Mallick and Smith (1998) develop stochastic searches via Markov chain Monte Carlo (MCMC) methods for Bayesian computation. Rather than greedily growing and then pruning a tree, these algorithms employ a number of move types to traverse the space of tree models. Move types include grow and prune steps, and a "change" step in which the rule at an interior node is changed. Chipman et. al. also propose a "swap" step in which the rules of a parent and child node are interchanged. In the algorithm, one of the four steps is utilized to generate a candidate tree by making a small random change to the current tree. The algorithm then either jumps to the new tree or remains at the current tree, depending on the ratio of posterior probabilities and transition proba-

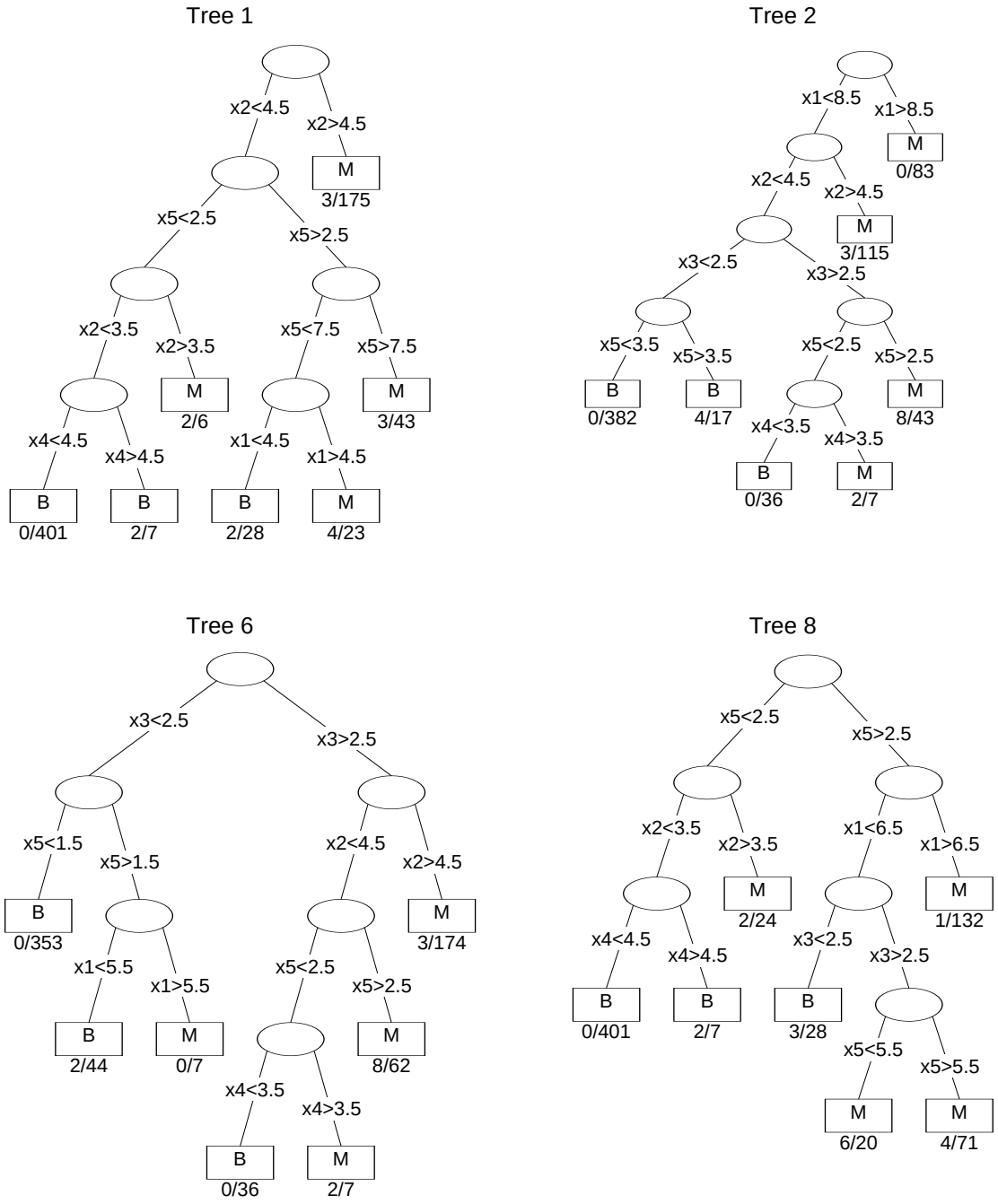


Figure 1: Four trees describing the breast cancer data. Trees 1,2,8 were identified by the Bayesian search, and 6 by bumping.

bilities between the current and candidate trees. Details of this Metropolis-Hastings algorithm are given in Chipman et. al. (1998).

Lutsko and Kuijpers (1994) develop a stochastic search algorithm based on simulated annealing. As with the Bayesian approaches, candidate trees are generated (using similar steps), and the resultant tree is either accepted or rejected. As the algorithm progresses, a temperature parameter is gradually decreased, making the algorithm less likely to jump to a candidate tree. Roughly speaking, the MCMC approach can be thought of as simulated annealing with a constant temperature (Geyer and Thompson, 1995).

# 3 Tree Metrics

The approach of the paper is to think of each tree as a point in a complex high-dimensional space, and cluster the trees according to some measure of proximity. Obviously, this space is much richer and more complicated than Euclidean space, and distances between trees can be measured in a number of fashions. To facilitate development of metrics, note that a tree can be identified by a finite set of parameters, and these parameters can be broadly divided in two groups: the tree itself and the parametric models in each terminal node. Referring to Figure 1, the tree parameters would include the splitting rules ($X_5 < 2.5, X_2 < 3.5$, etc) and the topology of the tree (the top node has two children, both of which are interior nodes, etc). The parametric model in each terminal node would be the probability of belonging to each response class (e.g. P(benign)=1/132, P(malignant)=131/132 for the rightmost terminal node). Metrics may be defined on either the tree or the terminal node parameters, or perhaps both. Below we propose three different metrics which capture different aspects of the tree.

Let $T_1, T_2$ be two trees with $b_1$ and $b_2$ terminal nodes. They have been trained using the same $n$ observations $(y_i, \mathbf{x}_i), i = 1, \ldots, n$.

For each observation $y_i$ we have an associated *fitted value* $\hat{y}_{ij}$ for tree $j$. The fitted value could be simply a mean or a class label. For a given tree and sample data with continuous response, the fitted value would be the average of all observations in that node. With a categorical response, the estimated class label for a node would be the class which had the highest sample proportion (assuming equal misclassification costs). The fitted values of the two trees can be used in a *fit metric*:

$$d(T_1, T_2) = \frac{1}{n} \sum_{i=1}^{n} m(\hat{y}_{i1}, \hat{y}_{i2}), \tag{1}$$

where $m$ is a metric on the fitted values. For regression trees with a continuous response, natural choices would be

$$m(y_1, y_2) = (y_1 - y_2)^2 \tag{2}$$

or $m(y_1, y_2) = |y_1 - y_2|$. For classification trees, $\hat{y}_{ij}$ might be the estimated class for observation $y_i$, in which case we could compare classifications by

$$m(y_1, y_2) = \begin{cases} 1 & \text{if } y_1 = y_2 \\ 0 & \text{otherwise} \end{cases} . \tag{3}$$

Metrics on the estimated class probabilities$(\hat{p}_{1j}, \ldots, \hat{p}_{cj})$, $j = 1, 2$ (for $c$ response classes) are also possible. A natural choice would be the Kullback-Leibler distance. Note that this and (2) have strong connections with likelihood functions for multinomial and normal data.

Rather than using the fitted values, a metric could be defined on the manner in which trees partition the predictor space. Trees which are very similar will place the same observations together and separate the same observations. Andrews (personal communication) suggests the following metric. Let $I_1(i, k)$ be 1 if $T_1$ places observations $i$ and $k$ in the same node and 0 otherwise. For a *partition metric*, we look at differences between $I$ for the two trees:

$$d(T_1, T_2) = \frac{\sum_{i>k} |I_1(i, k) - I_2(i, k)|}{\binom{n}{2}}. \tag{4}$$

The factor $\binom{n}{2}$ scales the metric to the range (0,1) with 0 indicating perfect agreement. A pair of observations contributes a positive amount to the distance only if one tree places the observations together and the other tree places them apart. The summation may be more efficiently calculated by noting that the argument will be identical for all observations that belong to the same terminal nodes. Thus a table of frequencies for nodes in $T_1$ and $T_2$ may be used to calculate (4).

Neither of these metrics account for the topology of the tree - they only use the observed responses and the partition defined by the terminal nodes. Shannon and Banks (1998) propose a *tree metric* which accounts for the manner in which the tree is constructed. This metric compares rules at nodes in the same position in the two trees. That is, if two plots are constructed on transparent paper so that nodes in the same position overlap and the plots are held up to the light, the metric counts the number of nodes at which the splitting rules are discrepant. The distance between trees is then a weighted sum of the discrepancies at each location:

$$d(T_1, T_2) = \sum_{r \in \text{nodes}(T_1, T_2)} \alpha_r m(\text{rule}(T_1, r), \text{rule}(T_2, r)) \tag{5}$$

The summation is over all node positions $r$ which are nonterminal in at least one tree. The metric $m$ compares the rules at two nodes; Shannon and Banks take $m$ to be 1 whenever the same variable is used (no matter what splitting rule is used within a variable), and 0 otherwise. Choosing all weights $\alpha_r = 1$ yields a count of the number of nodes at which the rules differ. If a node at a given location is nonterminal in one tree and either terminal or does not exist in the other tree, the nodes are considered different and are counted in (5).

# 4 An Example

In this section we use the breast cancer data (discussed briefly in the introduction) to illustrate an approach that may be used either interactively or in an automated fashion. Although illustrated with a specific example, the techniques used are applicable in general. In section 4.1, the data are described, and bumping and MCMC methods are used to generate a forest of trees. Multidimensional scaling is used in section 4.2 to visualize the distances between trees and compare the diversity of trees produced by bumping and Bayes methods. In section 4.3 we present the main part of our approach, in which a few representative trees are chosen from the forest. By considering trees one at a time, starting from the best fitting tree, we are able to characterize a forest as unimodal (good trees all similar) or multimodal (good trees form distinct clusters). A new graphic, the *added tree plot* is introduced to determine how many good trees are needed to cover the forest.

## 4.1 Tree generation

We now describe the data and the manner in which the forest was generated. The goal of the study is to classify breast tumors as benign or malignant so the response is binary. The predictor variables (listed in Table 1) consist of nine cellular characteristics each of which is measured on a 1-10 scale. Based on some preliminary analysis we used just five of the nine. Thus, each rule for splitting the observations in a particular node is of the form $x_i \leq c$ or $x_i > c$ where $c$ is one of the values 1.5,2.5,...9.5 and $i$ refers to one of the five predictors. Figure 1 displays four different trees that we have found to capture the pattern in the data. Note that the trees *are different*. For example, all four use a different variable to split on at the top node. How did we find these trees and how are they representative of the tree space?

Our first step was to generate a forest. Our goal was to obtain a set of trees that represent the variety of plausible tree models. We used both bumping and Bayesian stochastic search to find trees. An initial cross-validation analysis suggested that good trees should have between 5 and 10 bottom nodes. We implemented the bootstrap approach by resampling 250 times. For each resampled dataset, three trees of size 5, 6, and 7 are selected with cost complexity pruning. From the 750 trees produced we selected the best 20 trees out of those having 5 bottom nodes, the best 20 having 6, and the best 20 having 7, giving a total of 60 trees. We implemented the Bayesian search algorithm by restarting the chain 20 times. Each chain was run for 5000 iterations. From each of the 20 runs we kept the best tree found with 5 bottom nodes, the best having 6, and the best having 7 for a total of 60 trees. The "best" tree was defined to be the one having the largest integrated log likelihood (see Chipman et. al. 1998). Of the 60 trees kept we had 9 duplicates so 51 different tree were actually found. Combining these 51 with the 60 bootstrap trees we obtained 111 trees. All pairwise distances between trees were calculated using metrics (1), (4), (5). For fit metric (1), misclassification distance (3) was used.

| Variable | Code |
|---|---|
| Clump Thickness | $X_1$ |
| Uniformity of Cell Size | $X_2$ |
| Uniformity of Cell Shape | $X_3$ |
| Marginal Adhesion | |
| Single Epithelial Cell Size | $X_4$ |
| Bare Nuclei | $X_5$ |
| Bland Chromatin | |
| Normal Nucleoli | |
| Mitoses | |

Table 1: Variable names, breast cancer data. Variables used in this example are labeled $X_1 \ldots X_5$.

## 4.2 Visualizing the forest

We would like to "see" our forest of 111 trees in the tree space. Multidimensional scaling (MDS) produces a two-dimensional scatterplot in which each tree is represented by a point. The points are arranged so that the Euclidean distances between points are as close as possible to the original distances between trees. Many multivariate analysis texts discuss MDS; see for example Johnson and Wichern (1992).

Figure 2 presents MDS plots for the fit metric and the tree metric. Points corresponding to trees found by the Bayesian search are plotted with an "x" and trees found by the bootstrap are plotted with an "o". An immediate and interesting insight provided by the clustering of the

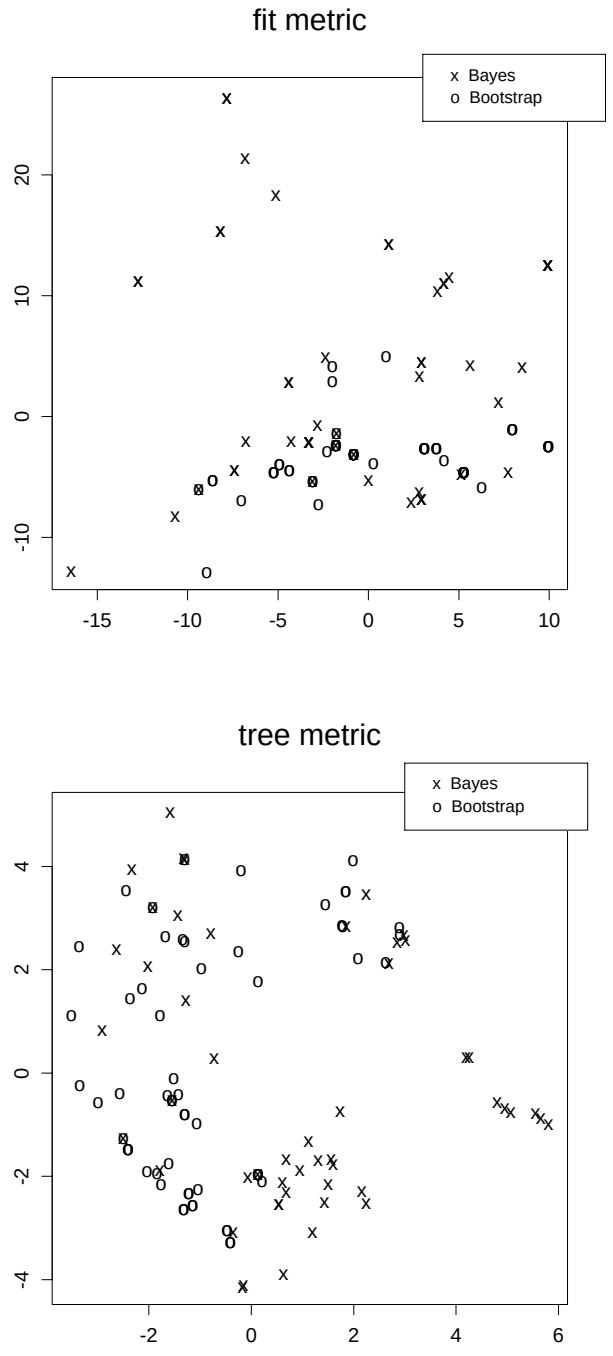


Figure 2: MDS plots for the breast cancer example. Fit (top plot) and tree (bottom plot) metrics are presented.

x's and o's is that the Bayes and bootstrap approaches have explored different parts of the tree space. For example in the tree metric plot there is a group of Bayes trees on the right hand side of the plot. Of course, the bootstrap approach may have found trees of this type if more resampling was done.

### 4.3 Picking representative trees

Our hope is that these 111 trees cover the tree space in that any "good tree" is reasonably close to one of the 111. The question now is: which of the 111 trees should be examined? A basic goal is to determine if there are different kinds of trees that seem to fit the data as opposed to all good trees being similar. In Bayesian terminology, we want to know if the posterior is multi-modal. We want to find a few trees which fit well *and* represent any variation in the kinds of trees that fit well. In this example the response is binary so we used (integrated) log likelihood as a criterion for good fit. The higher the log likelihood, the better the tree.

Figure 3 again plots the 111 trees using the MDS coordinates for the tree metric. The 13 trees with largest log likelihood are successively numbered (1 being the highest log likelihood). Throughout this example, we refer to trees by the rank of their corresponding log-likelihoods (i.e. 1-13). Thus, tree 1 is the "best tree" (as measured by log likelihood). In this example the forest is clearly multimodal. Four possible clusters of similar trees that fit well are given by {1,3,5,7,11}, {2,4,9}, {6,13}, and {8,10,12}.

By highlighting the most likely trees, the clusters come into sharper focus. Trees 1 and 2 (the two best trees) fall in different clusters so that we have found two different kinds of trees that both potentially fit the data. In order to represent the tree space with a small number of trees we choose a representative tree from each cluster. We chose trees 1, 2, 6, and 8 since they have the highest log likelihood among nearby trees. Recall that these four trees are displayed in Figure 1. As noted above, these trees are quite different. They may mean different things to the investigator with subject matter knowledge.

As we move down the list of trees from best to worst, we look to see if a new tree is different from the ones before. For example, in figure 3 we see that when we get to the eighth tree we have a tree which is in a different part of the tree space than the first seven. This will happen when the forest of trees is *multimodal* in that there are clusters of trees in different parts of the tree space that fit the data well. If we had a three dimensional plot of the log likelihood vs the two MDS coordinates we would see local maxima. As we "lower the bar", that is go down the list of trees ordered by log likelihood, different clusters corresponding to different local maxima become apparent. In contrast suppose there was really just one kind of tree that fit well. Then the forest would

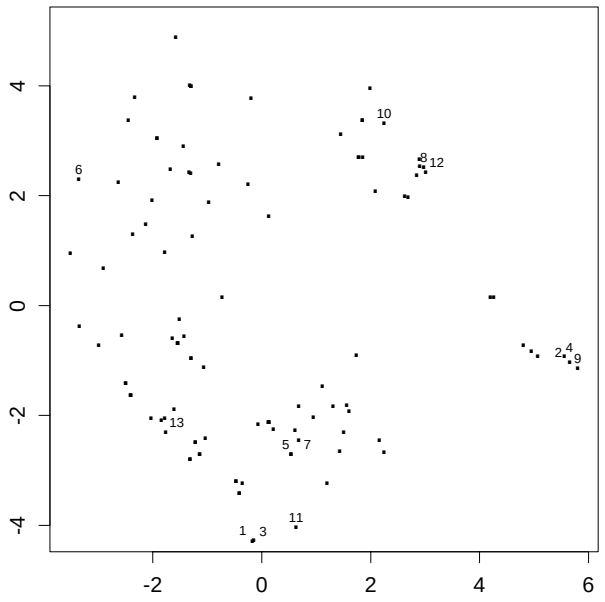


Figure 3: MDS plot of 111 trees using tree metric, with the most likely trees identified.

be unimodal. As we lower the bar, the new trees would spread out from the location of the best tree.

The core technique used above may be abstracted as follows:

1. Starting with the best tree, add trees one at a time until the space of trees is adequately represented. That is *"lower the bar"* until no new trees add diversity.

2. Cluster the trees identified in (1) and pick a representative tree from each cluster (either the most likely or most central).

While the MDS plot enables clear identification of clusters, it may distort the actual tree distances. As a useful alternative, we introduce the *added tree plot*, which uses the original distances and implements steps (1) and (2) above. The added tree plot in Figure 4 uses the original tree distances to assess the effect of adding new trees one at a time from left to right. For each index value on the horizontal axis, all distances between the new tree and all better trees are plotted on the vertical axis. For example, in the tree metric, the second best tree has a distance of 8 from the best tree (and is consequently quite different). The third best tree in this metric is quite close to one of the two best trees, and distant from the other. The vertical scale (for distance) in the added tree plot has range 0 to the maximum distance among all 111 trees, allowing us to see how diverse the best trees are relative to the whole forest. We can see that some distances are not accurately represented in the MDS plot. For example, in Figure 4 (with the tree metric), the nearest tree to 4 (out of 1-3) has a distance of 5, while in the MDS plot 2 and 4 nearly overlap.

These plots may be used to decide how many trees to include, i.e. how far to lower the bar. A tree should be added if it is far from other trees. In the added tree plot, this usually corresponds to a large minimum distance, meaning that no other tree is close. In the case of the tree metric, tree 7 is redundant, since it has zero distance to another tree. Tree 8 is different from 1-7, since its closest neighbour is a distance of 8. Based on only the added tree plot for the tree metric, using either the top 8 or the top 11 trees seems most reasonable. This agrees with the MDS plot, as we see that both plots identify the introduction of trees belonging to new clusters. Our choice of 13 trees perhaps includes too many, but the much worse error would be to include too few trees.

In some cases interesting trees may have large maximum distances, rather than large minimum distances. This means a tree is farther away from another tree than any other tree yet added. Such a tree extends the boundary of the forest, and the only trees which are close are also near the boundary.

The strategy in the previous two paragraphs may be extended to simultaneously look at several metrics. To select a cutoff, we would identify the largest index for which all successive trees add little diversity in any index. In this case, either eight or 11 trees is probably sufficient, although the 16th tree is not close to others in the partition metric.

Disagreements between metrics on individual trees can be quite informative. If a tree is similar to others in one metric and different in another, then only certain aspects of the tree are unique. Tree 8 is quite different from 1-7 in the tree metric, but not so different from these trees in the partition metric. This indicates that perhaps different rules are being used (or perhaps a different configuration of the same rules) to arrive at a partition that is not very different from the other trees.

Notice that in the added tree plot for the fit metric, none of the 20 trees considered comes close to the maximum distance of all 111 trees. Put another way, the fit metric does not discriminate among the most likely trees as much as the other two metrics. The fact that we

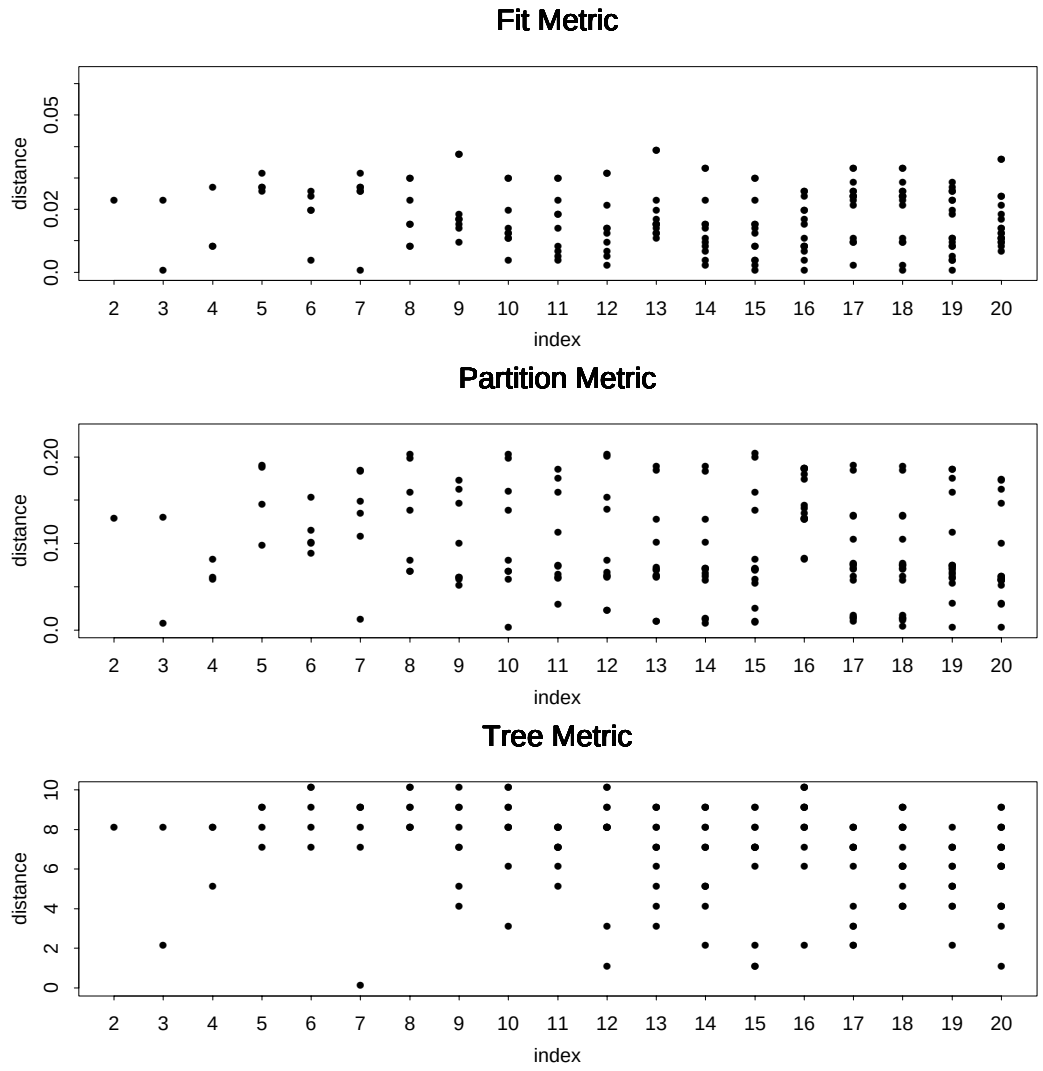


Figure 4: Added tree plots for the three metrics. For each index value, distances between the new tree and all better trees are plotted on the vertical axis. The vertical axes have maximum distance equal to the largest distance among all 111 trees.

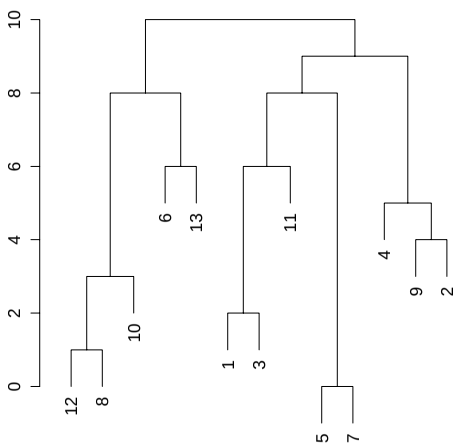


Figure 5: Dendogram for divisive hierarchical clustering of 11 trees with the tree metric.

are ordering trees by log likelihood implies that all trees with high log likelihood will probably be close in the fit metric. For this reason we feel that the fit metric is the least informative of the three considered.

Now that a set of good (and diverse) trees have been identified, they need to be clustered. The informal clustering based on MDS is probably sufficient for this example, since the clusters are so distinct. An alternative would be an automatic clustering procedure. Figure 5 shows the result of divisive hierarchical clustering (Kaufman and Rousseeuw 1990, implemented as `diana` in Splus) applied to the 13 selected trees. We see that the clustering basically identifies the same groups of trees which were visually evident in the plots based on multidimensional scaling. In this case the more automatic approach based on clustering gives the same results we got graphically using MDS.

# 5 Discussion

While the three metrics considered here are all useful, each has limitations. The fit metric lacks discriminative power, since all trees one considers will have reasonable fit to the data. If the same rules appear in different order, the tree metric will indicate that the trees are very different. Modifications of these metrics and other new metrics might address some of these issues. For example, before applying the tree metric, one might preprocess trees so that isomorphic trees are reduced to a common structure. Other metrics are also possible: One might construct a vector of summaries for each tree (eg size, bushiness, degree of use of each variable, etc.) and apply a metric to suitably normalized vectors. Another possibility is to regard each node as a sequence of rules combined by "ands", reorganize the rules to a common format, and apply a metric to the rules.

Additional techniques for clustering trees using these metrics are possible. The partition metric could be decomposed into a contribution for each observation, giving a measure of how "contentious" an individual observation is - that is, if two trees are different, what nodes do they disagree on most?

Specific to MCMC, these metrics may be used to aid in the running of the chain, and could provide useful (Bayesian) summaries. As Chipman et. al. (1998) discuss, running many chains is the most effective means to explore the space of trees. Metrics could be used to help decide how many chains to run and for how long. Using the metrics, we can examine the diversity of trees produced by each run (in a manner similar to the Bayes/bump comparison of Figure 5), and identify when a specific run is no longer exploring new territory. A more radical use of metrics might be to incorporate them in the construction of new candidate trees. If two distant trees have been identified (perhaps originating in different runs), then one might construct a chain that moved from one tree to the other, looking for interesting trees that exist between the two modes. Another application of clustering might be to calculate the posterior mass associated with clusters. Although calculating posterior probabilities of individual trees is problematic (Chipman et. al. 1998), these difficulties are alleviated by looking at mass associated with groups or neighbourhoods of trees.

Papers by other authors consider related issues. Shannon (1998) looks at predictive accuracy of trees identified as interesting using similar metrics. Hawkins and Musser (1998) use a forest of trees to learn what variables tend to occur together or apart in individual trees. Shannon and Banks (1998) propose the tree metric and use it to construct a single tree that is central to a forest. An important distinction is that they do not rank the trees according to their fit, a central element of our approach.

MCMC and other stochastic methods are being used increasingly to explore large and complicated model spaces. In situations where numerous models are indexed by a goodness criterion and a few representative models must be selected, the added tree plot of section 4.3 may have general application. In such general cases,

we propose that the plot be called an *added model plot*.

## Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council of Canada, NSF grant DMS 94.04408, Texas ARP grant 003658130, and research funding from the Graduate Schools of Business at the University of Chicago and the University of Texas at Austin. The authors would like to thank David Andrews, David Banks, Douglas Hawkins, Bret Musser, Sunil Rao, and William Shannon for stimulating discussions related to this work.

# References

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

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

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

Geyer, C. J. and Thompson, E. A. (1995) "Annealing Markov Chain Monte Carlo With Applications to Ancestral Inference", *Journal of the American Statistical Association*, 90, 909–920.

Hawkins, D.M. and Musser, B.J. (1998) "One Tree or a Forest? Alternative Dendrographics Models", *Proceedings of the 30th Symposium on the Interface.*

Johnson, R. A. and Wichern, D. W. (1992) *Applied Multivariate Statistical Analysis, 3rd edition*, Prentice Hall, Upper Saddle River.

Kauffman, L. and Rousseeuw, P. J. (1989) *Finding Groups in Data : An Introduction to Cluster Analysis*, Wiley, New York.

Lutsko, J. F. and Kuijpers, B. (1994) "Simulated Annealing in the Construction of Near-Optimal Decision Trees", in *Selecting Models from Data: AI and Statistics IV*, P. Cheeseman and R. W. Oldford, Eds., 453–462.

Quinlan, J. R. (1996) "Bagging, Boosting, and C4.5", *Proceedings of the Thirteenth National Conference on Artificial Intelligence AAAI '96).*

Shannon, W. (1998) "Averaging Classification Tree Models", *Proceedings of the 30th Symposium on the Interface.*

Shannon, W., Banks, D. (1998) "Combining Classification Trees using MLE', *Statistics in Medicine*, In Press.

Tibshirani, R., and Knight, K. (1995), "Model Search and Inference by Bootstrap 'Bumping' ", University of Toronto technical report.

Wolberg, W. H. and Mangasarian, O. L. (1990), "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, 87, 9193-9196.