


An Introduction to Random Forests (Self-Study)


There are several random forest algorithms. The approach described in this section is implemented by the HPFOREST procedure in SAS Enterprise Miner. A different algorithm is used by the RANDOMWOODS statement in PROC IMSTAT.

The SAS logo, consisting of the letters "sas" in a stylized font, with the tagline "THE POWER TO KNOW." to its right.

Forest

- An *ensemble model* is an aggregation of more than one model where the final prediction of the model is a combination of the predictions from the component models of the ensemble.
- A *forest model* is an ensemble of classification or regression trees.
- Forest models were developed to overcome the instability that a single classification or regression tree exhibits with minor perturbations of the training data.

95

The SAS logo, consisting of the letters "sas" in a stylized font, with the tagline "THE POWER TO KNOW." to its right.

Seeing the Forest through the Trees...

Trees in the forest differ from each other in two ways:


- Training data for a tree is a sample without replacement from all observations.
- Input variables considered for splitting a node are randomly selected from available inputs. Only the variable most associated with the target is split for that node.

96

The trees that make up a forest differ from each other in two ways:

- The training data for a tree is a sample without replacement from all observations that were originally training data for the forest.
- The input variables considered for splitting for any given node are selected randomly from all available inputs.

Among these variables, only the variable most associated with the target is used when forming a split. This means that each tree is created on a sample of the inputs and from a sample of observations. This process, repeated many times, creates a more stable model than a single tree. The reason for using a sample of the data to construct each tree is because when less than all available observations are used, the generalization error is often improved. A different sample is taken for each tree.

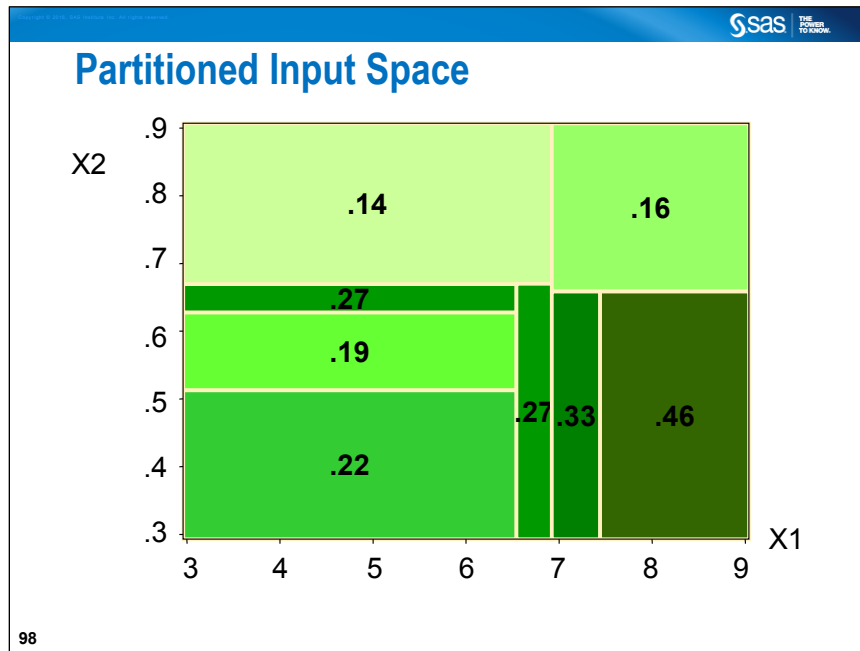
			
Leaves = Boolean Rules			
If $X1 \in \{values\}$ and $X2 \in \{values\}$, then $\hat{Y}=value$.			
Leaf	X1	X2	Predicted Y
1	<6.5	<.51	.22
2	<6.5	[.51, .63)	.19
3	<6.5	[.63, .67)	.27
4	[6.5, 6.9)	<.67	.27
5	<6.9	\geq .67	.14
6	[6.9, 7.4)	<.66	.33
7	\geq 7.4	<.66	.46
8	\geq 6.9	\geq .66	.16

97

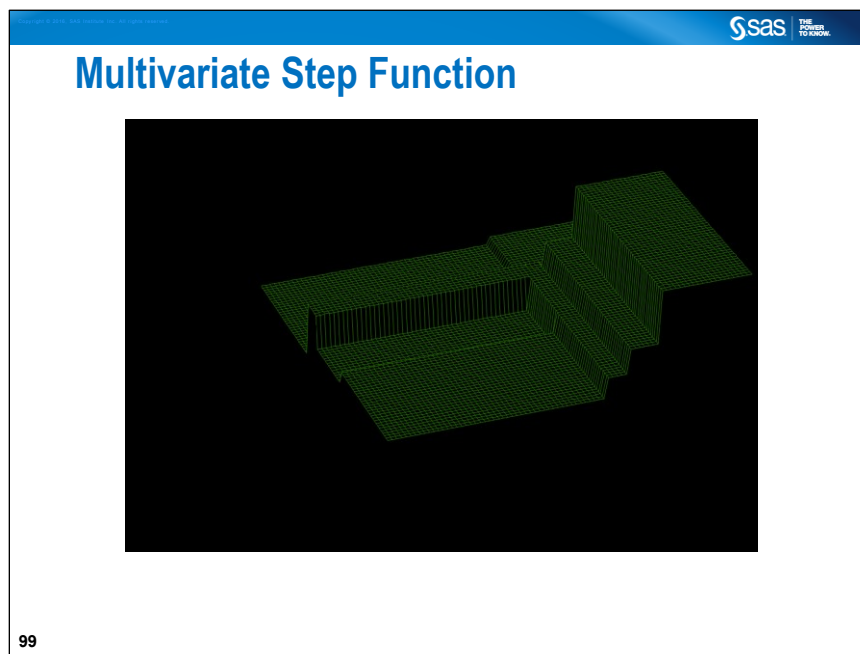
The path to each leaf can be expressed as a Boolean rule. The rules take this form:

If the inputs $\in \{region\ of\ the\ input\ space\}$, then the predicted value = *value*.

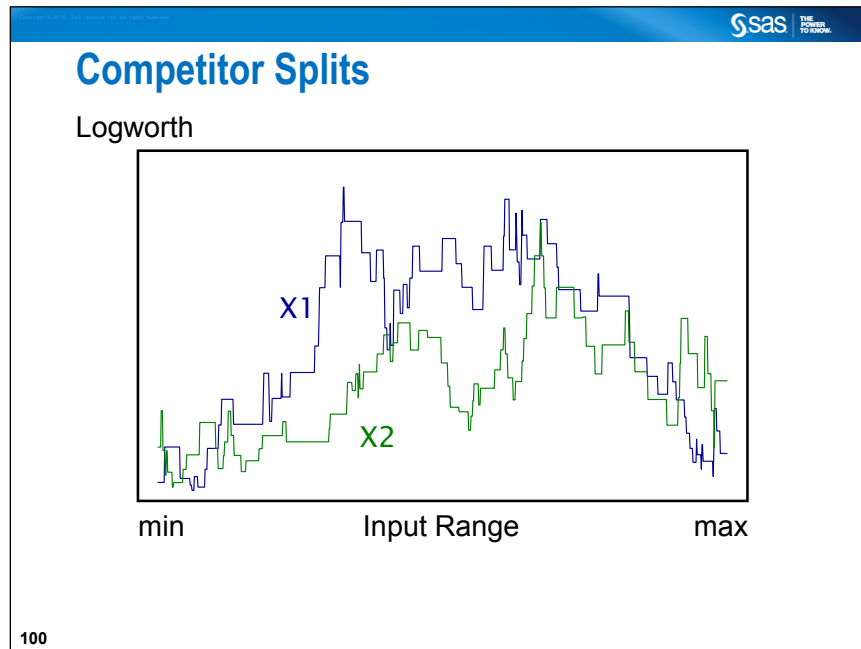
The regions of the input space are determined by the split values. For interval-scaled inputs, the boundaries of the regions are perpendicular to the split variables. Consequently, the regions are intersections of subspaces defined by a single splitting variable.



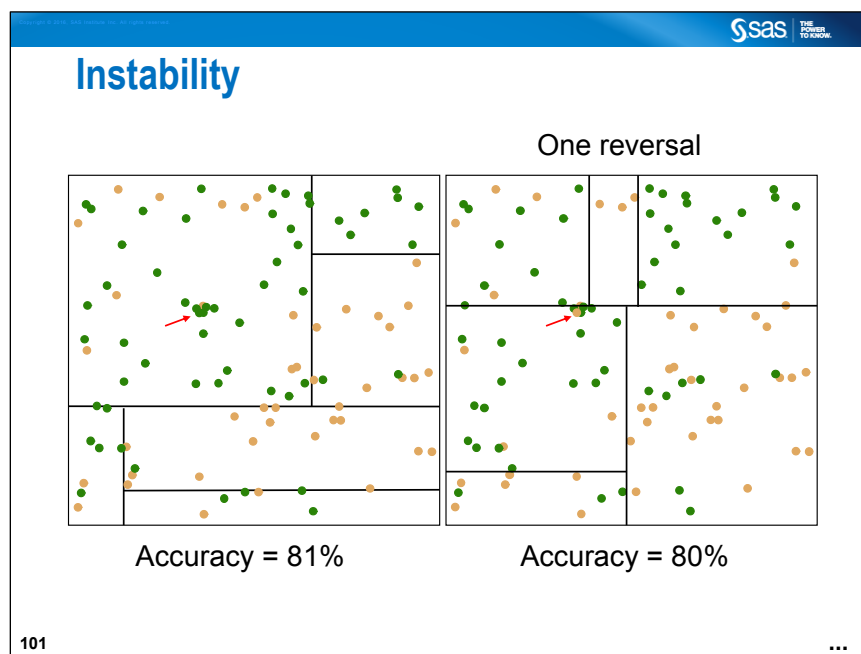
The leaves of the decision tree partition the input space into rectilinear regions. The predicted target has a different constant value in each partition. Consequently, the fitted model is a multivariate step function.



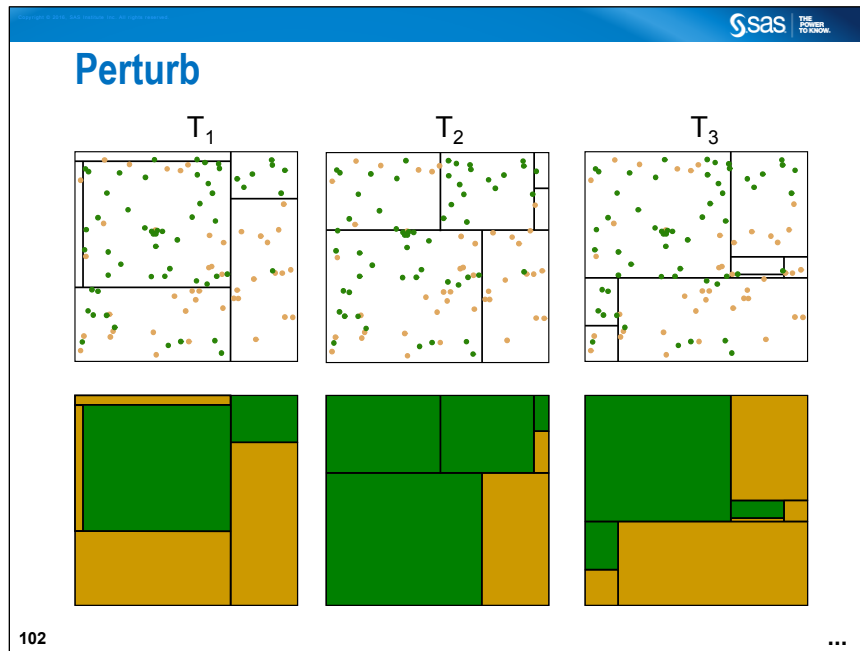
The surface is a piecewise constant and not joined continuously at the boundaries. A step function is highly flexible. It is capable of modeling nonlinear trends.



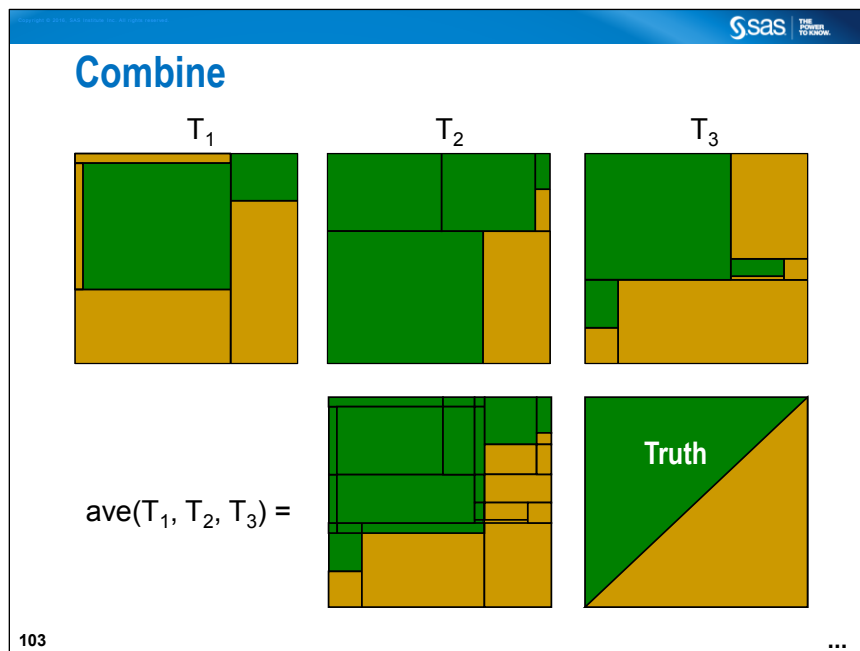
Decision trees are unstable models. That is, small changes in the training data can cause large changes in the topology of the tree. However, the overall performance of the tree remains stable (Breiman et al. 1984). The instability results from the large number of univariate splits considered and the fragmentation of the data. At each split, there are typically a number of splits on the same and different inputs that give similar performance (competitor splits). A small change in the data can easily result in a different split being chosen. This in turn produces different subsets in the child nodes. The changes in the data are even larger in the child nodes. The changes continue to cascade down the tree.



In the above example, changing the class label of one case resulted in a completely different tree with nearly the same accuracy.



Methods have been devised to take advantage of the instability of trees to create models that are more powerful. *Perturb and combine* (P & C) methods generate multiple models by manipulating the distribution of the data or altering the construction method and then averaging the results (Breiman 1998). Any unstable modeling method can be used, but trees are most often chosen because of their speed and flexibility.



An ensemble model is the combination of multiple models. The combinations can be formed by

- voting on the classifications
- weighted voting where some models have more weight
- averaging (weighted or unweighted) the predicted values.

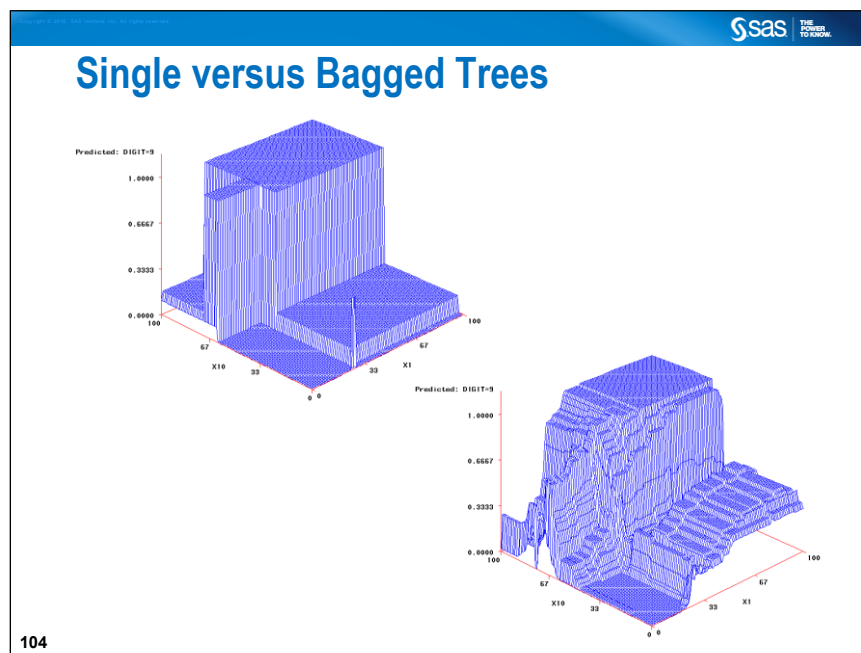
Ensemble methods are a very active area of research in the fields of machine learning and statistics. Many other P & C methods were devised.

The attractiveness of P & C methods is their improved performance over single models. Bauer and Kohavi (1999) demonstrated the superiority of P & C methods with extensive experimentation. One reason why simple P & C methods give improved performance is variance reduction. If the base models have low bias and high variance, then averaging decreases the variance. In contrast, combining stable models can negatively affect performance. The reasons why adaptive P & C methods work go beyond simple variance reduction and are the topic of much current research (Breiman 1998). Graphical explanations show that ensembles of trees have decision boundaries of much finer resolution than would be possible with a single tree (Rao and Potts 1997).

A new case is scored by running it down the multiple trees and averaging the results. Multiple models need to be stored and processed. The simple interpretation of a single tree is lost.

Bagging goes a long way toward making a silk purse out of a sow's ear; especially if the sow's ear is twitchy. ...What one loses, with the trees, is a simple and interpretable structure. What one gains is increased accuracy.

— Breiman (1996)



To smooth the all-or-none bins of a single decisions tree, bagging smooths the prediction surface.

Bagging (bootstrap aggregation) is a P & C method that generally works as follows (Breiman 1996):

1. Draw B bootstrap samples.

A bootstrap sample is a random sample of size n drawn from the empirical distribution of a sample of size N . That is, the training data is resampled with replacement. Some of the cases are omitted from the sample, and some cases are represented more than once.


2. Build a tree on each bootstrap sample.

Pruning can be counterproductive (Bauer and Kohavi 1999). Large trees with low bias and high variance are ideal.

3. Vote or average.

For classification problems, take the mean of the posterior probabilities or take the plurality vote of the predicted class. Bauer and Kohavi (1999) found that averaging the posterior probabilities gave slightly better performance than voting. Take a mean of the predicted values for regression.

Breiman (1996) used 50 bootstrap replicates for classification and 25 for regression and for averaging the posterior probabilities. Bauer and Kohavi (1999) used 25 replicates for both voting and averaging.


The SAS logo, consisting of the letters "sas" in a stylized font, with the tagline "The Power To Know." to its right.

Forest Algorithm

- *Bagging* is the term for averaging many trees grown on bootstrap samples of the rows of training data. All columns are considered for splitting at every step.
- The forest algorithm does sampling of the rows **and** sampling of the columns at each step.
- The forest algorithm perturbs the training data more than the bagging algorithm. Thus, it produces more variation among the trees in the ensemble.
- Ensembles of a more diverse set of trees often lead to improved predictive accuracy.

105

In a forest, rather than taking bootstrap samples of only the rows, variables are also randomly sampled. This results in a forest, consisting of trees that use different combinations of rows and variables to determine splits. This additional perturbation (beyond bagging) leads to greater diversity in the trees, and better predictive accuracy.



A New Term Is Needed

- The out-of-bag sample refers to the training data that is excluded during the construction of an individual tree.
- Observations in the training data that are used to construct an individual tree are the bagged sample.
- Some model assessments, such as the iteration plots, are computed using the out-of-bag sample as well as all the training data.

106

A decision tree in a forest trains on new training data that are derived from the original training data. Training different trees with different training data reduces the correlation of the predictions of the trees, which in turn should improve the predictions of the forest. The training data for an individual tree exclude some of the available data. The data that are withheld from training are called the *out-of-bag sample*. Observations in the training sample are called the *bagged* observations, and the training data for a specific decision are called the *bagged data*. For each individual tree, the out-of-bag sample is used to form predictions. These predictions are more reliable than those from training data.

Model assessment measures, such as misclassification rates, average squared error, and iteration plots, are constructed on both the entire training data set as well as the out-of-bag sample.


Variable Importance in a Forest

sas THE POWER TO KNOW.

Gini Impurity


$$1 - \sum_{j=1}^r p_j^2 = 2 \sum_{j < k} p_j p_k$$

high diversity, low purity



Pr(interspecific encounter) = $1 - 2(3/8)^2 - 2(1/8)^2 = .69$

low diversity, high purity



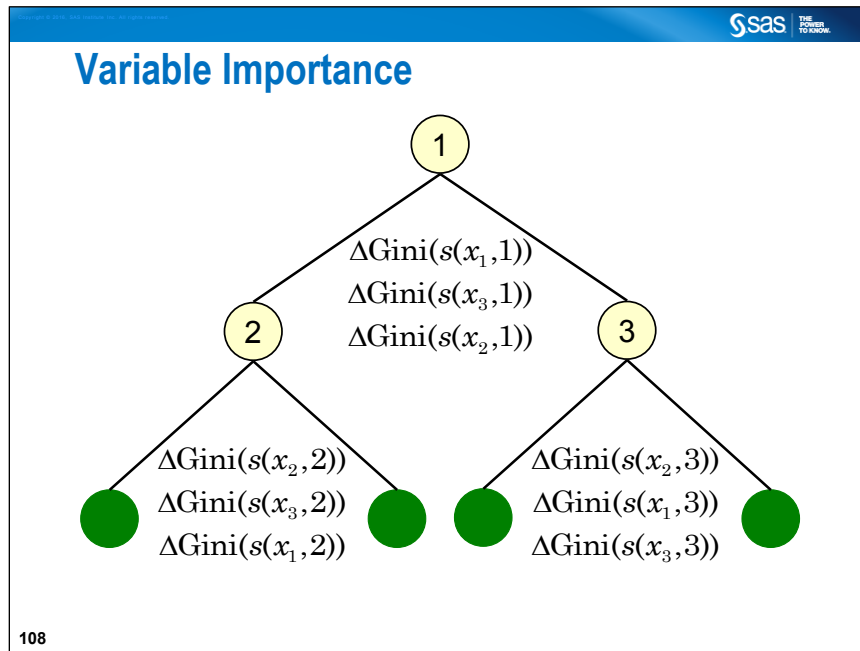
Pr(interspecific encounter) = $1 - (6/7)^2 - (1/7)^2 = .24$

107

The Gini index is a measure of variability for categorical data (developed by the eminent Italian statistician Corrado Gini in 1912). The Gini index can be used as a measure of node impurity where p_1, p_2, \dots are the relative frequencies of each target class in a node. The Δ Gini splitting criteria was proposed by Breiman et al. (1984).

The Gini index can be interpreted as the probability that any two elements of a multi-set, chosen at random (with replacement), are different. A pure node has a Gini index of 0. As the number of evenly distributed classes increases, the Gini index approaches 1.

In mathematical ecology, the Gini index is known as *Simpson's diversity index*. In cryptanalysis, it is 1 minus the *repeat rate* (good discussion of Patil and Taillie, 1982).




Breiman et al (1984) devised a measure of variable importance for trees. It can be particularly useful for tree interpretation.

Let $s(x_j, t)$ be a surrogate split (including the primary split) at the t th internal node using the j th input.

Importance is a weighted average of the reduction in impurity for the surrogate splits using the j th input across all the internal nodes in the tree. The weights are the node sizes.

$$\text{Importance}(x_j) = \sum_{t=1}^T \frac{n_t}{n} \Delta(i)(s(x_j), t)$$

In a decision tree, variable importance can be calculated similarly to Breiman et al (1984). The variable importance measure is scaled to be between 0 and 1 by dividing by the maximum importance. Thus, larger values indicate greater importance. Variables that do not appear in any primary or saved surrogate splits have 0 importance.

sas
THE POWER TO KNOW.

Summary Points

- Trees automatically handle missing values and variable reduction. Therefore, the input data requires less preparation.
- Forests tend to give better prediction than any specific tree, and often outperform other classes of models.
- Forests are challenging to interpret, but they can be considered an “ideal” model for other models to be compared against.

109

Variable importance measures provide a measure of interpretation for forests.