Introduction

Trees and Random Forests

To begin our discussion of trees and random forests, we will first consider the following example using data from a dendrologic study of five orange trees. This study measured two variables for each tree: the age of the tree (recorded in days) and the circumference of the trunk (in cm). These are called, in general terms, the variables recorded by the study. In the dataset below, each column represents a variable and each row represents one set of measurements. There are 2 columns (age and circumference) and 35 rows. The first six rows are displayed in table 1.

Table 1: The first six rows of the Orange data set

age	circumference
118	30
484	58
664	87
1004	115
1231	120
1372	142

Let's pretend we are interested in the following question: knowing only the circumference of the orange tree, can we predict the age of the tree? This question can be translated into a formula, or a guess at how the relation between the two variables functions. We often refer to formulas using the following notation:

$$Age \sim Circumference$$

In this formula, circumference is the predictor and age is the response. Suppose we expanded the study to include the height of the orange trees at various stages of development. Now we can consider both the circumference and the height of the tree when we make our predictions of the age. When we have multiple predictors, we add them to the notation in the following way:[1]

$$Age \sim Circumference, Height$$

Returning to the original orange tree data set, we can begin to assess the structure in the data set by plotting the data and observing the relationship between the variables in figure 1.

[1]Often the notation for multiple predictors is written $Y \sim X + W$ but this assumes an additive, linear relationship between the predictors and the response. This assumption is unnecessary for tree-based models so the notation, $Y \sim X, W$ is used.

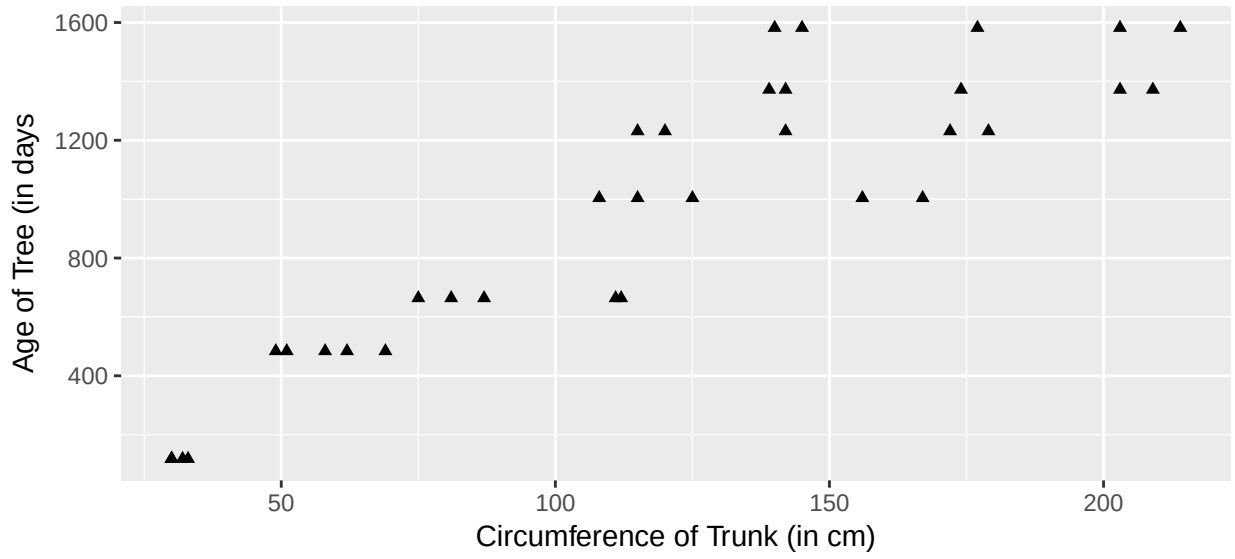

Figure 1: The relationship between age and circumference of the trunk of orange trees.

As can be seen in figure 1, generally older trees have thicker trunks, and it seems like we are not wrong to suspect that circumference is a good predictor of age. As the data could be reasonably represented by a straight line, we can say that the relationship between trunk circumference and tree age is roughly linear. To create our predictions of age, we fit the formula $Age \sim Circumference$ to a model. A predictive model is, put simply, a systematic way to make our predictions. The most common type of model is the linear model, which creates predictions from a line through the data.

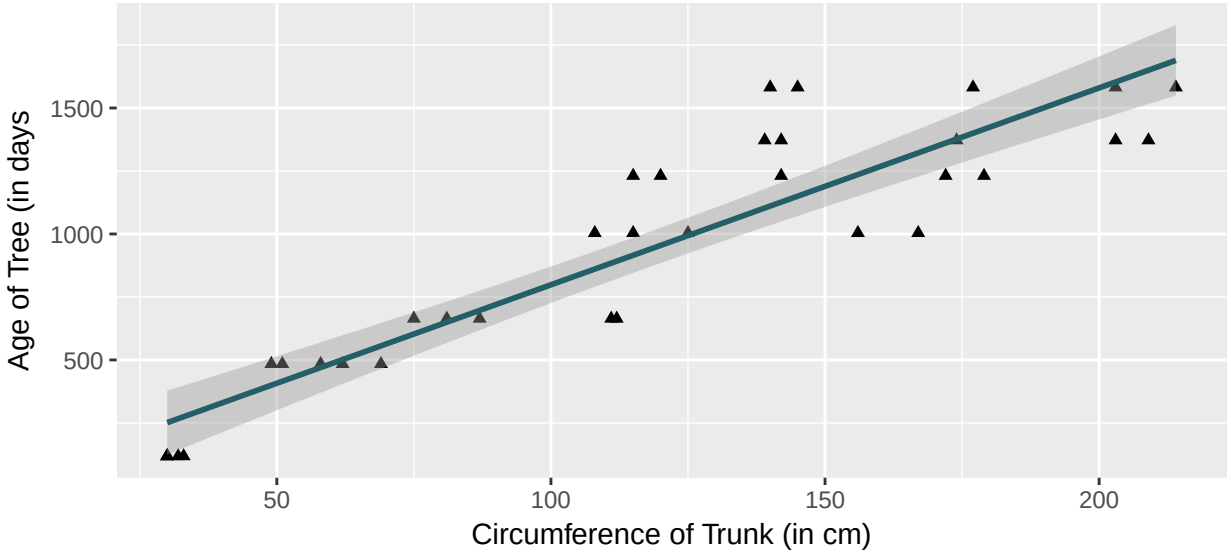

Figure 2: A linear model representing age ~ trunk circumference in orange trees. The shaded area represents a 95% confidence interval around this line.

As can be guessed from figure 2, the linear model works better on certain data than others. The linear model necessitates several assumptions that may not always be appropriate. We'll return to a brief discussion of the assumptions of the linear model later in this chapter.

This paper discusses at length tree-based models. A tree for the formula $age \sim circumference$ is similar to the linear model in that it presents a systematic way to make predictions, but the two differ in that the tree

is not linear in any fashion. In fact, we can compare the differences between the two models by comparing figures 2 and 3.

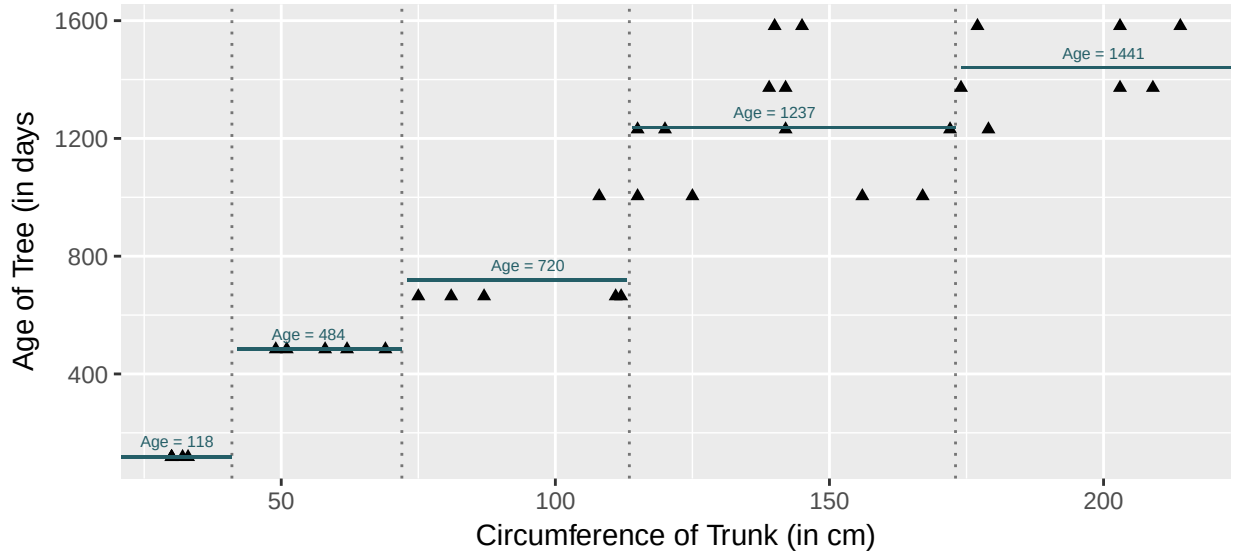

Figure 3: A tree modeling the formula age $\sim$ trunk circumference first creates partitions on the predictor, seen as vertical lines, and then predicts the value of the response within that partition, seen as text.

When using the linear model, we make predictions in the following way: given a value of X (circumference) the corresponding value of Y (age) on the line is our prediction. The predictions from the tree are gathered similarly: given a value of X (circumference), our prediction is the average value of Y (age) within the partition that X falls into.[2]

Tree methods get their name from a common way of representing them in higher dimensions, when there is more than one predictor. Figure 4 shows this method. In this case, given a new value for circumference, one would start their predictions at the top of the tree and, depending on the value of circumference and the instructions at each intersection or split, one would fall down branch by branch before landing on a prediction for age. If the new value for circumference is less than the value stated at the split, we would move down to the branch on the left. If circumference was more than the value for circumference at the split, we would move down to the branch on the right.

[2]While any curve can be approximated in a step-wise manner when the number of steps approaches ∞, a tree model does not converge to the linear model as the number of splits approaches the number of observations, even when the data is linear.

Table 2: The first six rows of the orange trees data set are repeated here (left) with row numbers. The table on the right represent a bootstrapped sample of the original data (left)

age	circumference	row
118	30	1
484	58	2
664	87	3
1004	115	4
1231	120	5
1372	142	6

	age	circumference	row
2	484	58	2
3	664	87	3
4	1004	115	4
6	1372	142	6
2.1	484	58	2
6.1	1372	142	6

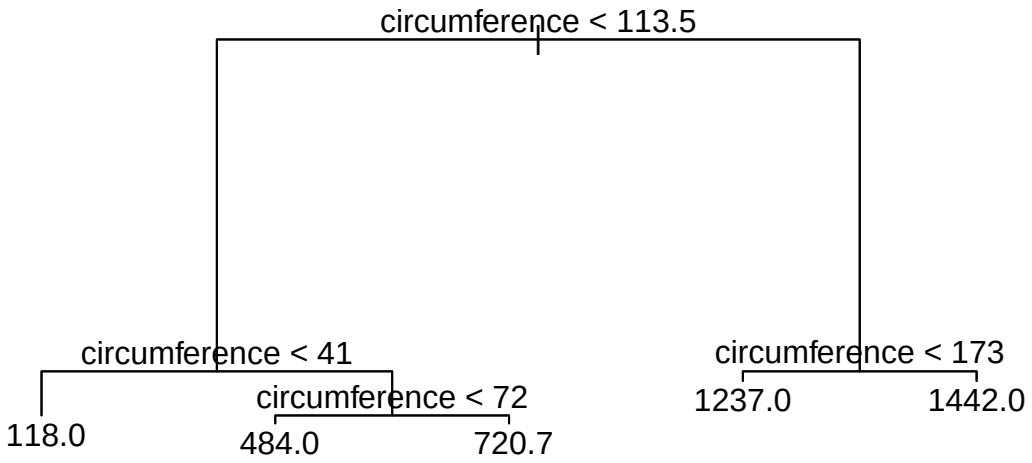

Figure 4: A tree representing age $\sim$ trunk circumference in orange trees.

Mathematical notation for trees

Given a data set, D, and a formula $Y \sim X$, where Y and X are columns in D, a tree, T, is a stepwise function from X to Y. X can also be defined a a set of columns in D. Recall, that X is the set of our predictors and Y is our response. In example in section 1.1, the orange trees data is the data set, the ages of the trees make up our response, and the circumferences are our predictor.

$$T : X \to Y$$

T is created using some sample of the rows of D. This sample is called the *training* set. The predictions or the image of T, are discrete values in the range of Y. We test the prediction accuracy of T using the rows of D that were not in the training set. This is called the *test* set. Often with trees, we sample D by *bootstrapping* the data set. "Bootstrapping" means sampling with replacement. This allows us to have large test and training sets without starting with an extra large data set D. This is valuable because it is more difficult to create good predictive models on small amounts of data relative to the entire population. The orange trees data set is an excelent candidate for bootstrapping, as there are only 35 records from 5 trees in the data set when there are presumably thousands of orange trees in the world at large. Table 2 demonstrates a bootstrapped sample of the first six rows of the orange trees data set from table 1.

There are many ways to test prediction accuracy of a model, but we will be considering the following two in this paper: MSE and RSS.[^3] MSE is the mean squared error and is defined as:

$$MSE(T, Y, X) = \frac{1}{n} \sum_{1}^{n} (Y - T(X))^2$$

Where n is the number of rows in X and Y. RSS is the residual sum of squares error and is defined as:

$$RSS(T, X, Y) = \sum_{1}^{n} (Y - T(X))^2$$

Often MSE and RSS are calculated on the test set to create an estimate of how well the model will be able to predict the outcome of a wholly new observation. Then they are refered to as MSE_{test} and RSS_{test}, respectively.

A brief history of trees

Trees are a convenient way to represent data and assist in decision making. Morgan and Sonquist (1963) derived a way for constructing trees motivated by the specific characteristics of data collected from interviews and surveys. The first difficulty in analyzing this data was that data collected from surveys is mostly categorical, where the observation is that the participant is a member of some discrete group. Some common categorical variables are gender, ethnicity, and education level. Numeric variables, like age, height, and weight are, in general, much easier to work with. On top of this, the data sets Morgan and Sonquist dealt with had few participants (rows) and many variables (columns). To add to their difficulties, there was reason to believe that there were lurking errors in the data that would be hard to identify and quantify. Lastly, many of the predictors were correlated. Morgan and Sonquist doubted that the additive assumptions of many models would be appropriate for this data. They noted that while many statistical methods would have difficulty accurately parsing this data, a clever researcher with quite a lot of time could create a suitable model simply by grouping values of the predictors and predicting that the response corresponding to these values would be an average of the observed responses given the grouped conditions. Their formalization of this procedure in terms of "decision rules" laid the groundwork for future research on decision trees. See figure 3 for a visualization of this process.

Later researchers proposed new methods for creating trees that improved upon the Morgan and Sonquist model. Leo Breiman et al (1984) proposed an algorithm called CART (Classification And Regression Trees) to fit trees on various types of data. Torsten Hothorn, Kurt Hornik and Achim Zeileis argue in their 2006 paper, *Unbiased Recursive Partitioning: A Conditional Inference Framework*, CART has a selection bias toward variables with either missing values or a great number of possible splits. This bias can affect the interpretability of all tree models fit using this method. As an alternative to CART and other algorithms, Hothorn et al. propose a new method: conditional inference trees. The conditional inference trees algorithm is similar to CART in many ways, but a thourough description is beyond the scope of this paper.

 In the 1984 textbook *Classification and Regression Trees*, Breiman, Friedman, Olshen, and Stone described their method for creating, pruning, and testing regression trees. There are essentially three steps: one, decide on a variable to split over, two, partition that variable space in two distinct partitions, and three, set our initial predictions for each partition to be mean value of the response according to the observed responses corresponding to the values in the partitions. Recursively, this process is repeated for each new partition until some stopping condition is reached. This is a top down, greedy algorithm that functions by creating as large a tree as possible [@bibCART].

Random Forests are generated by fitting a large number of trees, each on a boosted sample of the data. The crucial difference, however, between the trees in CART and the trees in a random forest, is that at each node in a random forest, only a subset of the predictors are considered as candidates for possible splits. This decorrelates each tree from its neighbors, and limits variability of the whole forest [@bibISL].

There is a limit to the predictive capabilities of a single tree as they suffer from high variance. These are further explored in chapter 2. To alleviate this, aggregate methods called forests are often used instead. They function by enlisting the help of many trees, and then by aggregating the responses over all of them. The two most common types of forests are bagged and random forests.

Mathematical notation for bagged and random forests

Given a data set, D, and a formula $Y \sim X$, where Y and X are columns in D, a forest, R, collection of trees from X to Y. As R is a collection of trees from X to Y, it is also a stepwise function from X to Y.

$$R : X \to Y$$

The trees in a forest are fit using bootstrapped samples of the original data set D. The bootstrapped training set for a tree T is B_t and the corresponding test set is $\bar{B}_t$. These test and training sets are called the *in bag* set and the *out of bag* set. When the MSE or RSS is calculated for a tree in a forest, it is done on the out of bag set for that tree and called the OOB (out of bag) error. The predictive accuracy of the random forest is the average of the out of bag error across all the trees in the forest. The predictions for the random forest, $R(X)$, are the average prediction of each tree on X.

These are *bagged forests*. *Random forests* are a variation on bagged forests, and the main focus of this paper. There is no difference between bagged and random forests when there is only one predictor in the data set, but, in cases with more than one predictor, the trees are generated in a slightly different way. Usually, to make a split all the predictors are considered. (Recall that splits are rules on certain predictors; rules like "if circumference is less than 113.5 go left, if not go right." The statement "circumference $<$ 113.5" from figure 4 is a split at 113.5 on the variable circumference). If we had a larger version of the orange tree data set that included the heights of each tree, then the tree would always consider both height and circumference as possible candidates for splits. In a random forest on this expanded orange tree data set, it is possible that only one predictor, height or circumference, would be considered at a time. The splitting procedure is discussed at length in chapter 3 and this property of random forests is discussed in chapter 4. For now only a cursory understanding is needed.

Inferential vs Descriptive Statistics

In the earlier sections, we focused on building predictive models, but this paper hopes to use tree-based methods beyond this context. The linear model is a mainstay in social science because it allows for easy and interpretable statistical inference. Return to the orange tree example from section 1.1. The linear model gives us a line with which we can make predictions, but it also gives estimated coefficients and conducts hypothesis tests on the values of the coefficients.

Table 3: Estimated linear coefficients, error, and p-values from the model fit in section 1.1 on the orange tree dataset

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16.603609	78.1406182	0.2124837	0.8330368
circumference	7.815998	0.6058806	12.9002281	0.0000000

This table provides evidence that not only is trunk circumference a good predictor of age, the relationship between them is the equation of the line:

$$Age = 7.81 \cdot Circumference + 16.6$$

Roughly, for every 1 cm of trunk growth, we would expect the tree to be 7.81 days older. Not only are we provided with a way to describe the relationship between age and circumference, we have conducted statistical tests to make reasonably sure that our estimates, 7.81 and 16.6 in the equation above, are not zero. Inferential claims about the nature of tree age and trunk growth are possible here.

It is important to note the difference between inferential and descriptive statistics. Descriptive statistics

describe the data at hand without making any reference to a larger data generating system that they come from. It follows that inferential statistics then make claims about the data generating system given the data. The model in figure 4 could be used to make descriptive claims about the orange tree data. For example, given the data we have, we expect a sapling with a trunk circumference less than 41 cm to be 118 days old. However, trees are variable; they are very sensitive to changes in the data set. It's entirely possible that if we fit this tree on a new sample of the data, the predictions would change. See chapter 2 for more discussion on the variability of trees. Our claims about the relation between circumference and age in young orange trees can only be descriptive as we have not taken into account the variability in gathering them. This paper's aim is to describe a process of making inferential claims using trees and random forests that employs permutation tests.

As stated in the introduction of the *Chronicle of Permutations Statistical Methods* by KJ Berry et al. 2014, there are two models of statistical inference. One is the population model, where we assume that the data was randomly sampled from one (or more) populations. Under this model, we assume that the data generated follows some known distribution. "Under the population model, the level of statistical significance that results from applying a statistical test to the results of an experiment or a survey corresponds to the frequency with which the null hypothesis would be rejected in repeated random samplings from the same specified population(s)" (Berry et al, 2014).

The permutation family of methods, on the other hand, only assumes that the observed result was caused by experimental variability. The test statistics are calculated for the observed data, then the data is permuted a number of times. The statistic is calculated after each permutation to derive a distribution of possible values under some null hypothesis. Then the original test statistic is tested against this distribution. If the observed value is exceptionally rare, then there is evidence that our observation did not come from that distribution.

Inference on Random Forests

The Problem

Random forests create models with great predictive, but poor inferential capabilities. After Morgan and Sonquist's initial development of decision trees, trees quickly moved to the domain of machine learning and away from statistics. Researchers focused on bettering predictions and improving run times and less on the statistics behind them. In a single tree, descriptive claims may be simple to make, but it is much more difficult to describe the behavior of the whole forest. Inferential statistics with random forests generally falls behind the predictions in importance. This has limited the applications of random forests in certain fields, as to many the question of "why" the data is the way it is is more important than building predictions. There are several means of performing descriptive statistics with random forests that could be interpreted incorrectly as attempting to answer this but without a statistically backed method for performing inference, the use of random forest is limited to prediction-only settings.

Proposed solutions to this problem

Variable importance could be the tree-based analogue to the coefficients of the linear model, in that the variable importance for the predictor X_i in the model for $Y \sim X_1, ..., X_p$ is the amount of predictive accuracy due to X_i. Breiman proposed a method of permuted variable importance in his paper *Statistical Modeling: The Two Cultures* to answer this problem. Their method compares the variable importance for each variable in a tree-wise manner. For each tree, the permuted variable importance of the variable X_j is:

$$VI^t(X_j) = \frac{\sum_{i \in |\bar{B}_t|}(y - \hat{y})^2}{|\bar{B}_t|} - \frac{\sum_{i \in |\bar{B}_t^*}(y - \hat{y}^*)^2}{|\bar{B}_t^*|}$$

Where $\bar{B}^t$ is the out of bag sample for tree t, $|B|$ is the number of observations in that sample, $\bar{B}_p^t$ is with X_j permuted, $\hat{y}$ is the predicted outcome, and $\hat{y}^*$ is the predicted outcomes after variable X_j has been permuted. This value is averaged over all the trees. It is important to note that if the variable X_j is not split on in the tree t, the tree-wise variable importance will be 0.

Strobl et al from the University of Munich criticize this method in their 2008 technical report *Danger: High Power! – Exploring the Statistical Properties of a Test for Random Forest Variable Importance*. First, this method has the downside of increasing power with increasing numbers of trees in the forest. This is a more or less arbitrary parameter which we would hope would not affect our importance estimates. Second, the null hypothesis under Breiman and Cutler's strategy is that the variable importance VI for any variable X_j is not equal to zero given Y, the response. Because random forests are most often used in situations with multicolinearity that would make other methods like the linear model difficult, Strobl argues that any variable importance measure worth its salt should not be misled by correlation within the predictors.

The researchers at the University of Munich published a fully fleshed response to the Breiman and Cutler method in 2008, titled *Conditional Variable Importance for Random Forests* that addresses these issues. Strobl et al propose restructuring the Breiman and Cutler algorithm to account for conditional dependence among the predictors. The null hypothesis is that $VI_\beta(X_j) = 0$ given the predictor Y *and all other predictors* $X_1, ..X_n$. This accounts for interactions between X_j and the other predictors, while preserving the relationship between Y and the remaining predictors.

This paper aims to provide a response to this method. The partitions are made from the random forest corresponding to the formula of $Y \sim X_1, ..., X_n$ instead of a model of $X_j \sim X_1, ..., X_n$. This ignores the common situation where if the predictors are correlated enough, then they act as stand ins for each other, so that if one variable is heavily influential in a certain tree at predicting Y, the other variable will be forgotten altogether.