# Introduction

## Trees and Random Forests

### Trees: An Example

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 things 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, each column represents a variable and each time each tree's measurements were taken is one row. There are 2 columns, age and circumference, and 35 rows. In table 1, the first six rows are displayed.

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 is called 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 our investigation into the validity of the formula 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.
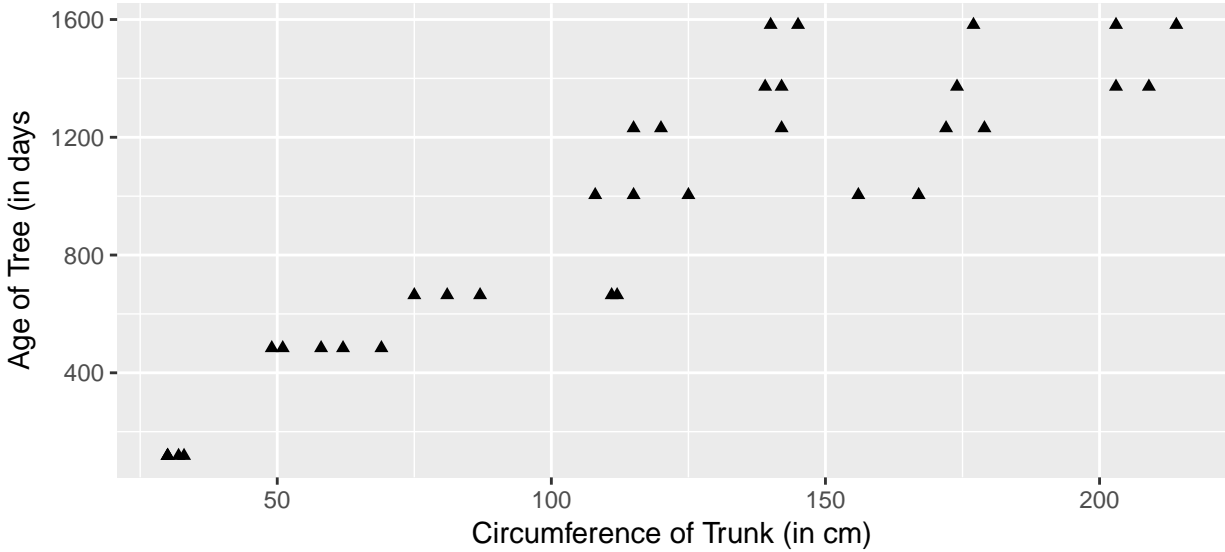
Figure 1: The relationship between age and circumference of the trunk of orange trees.

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 circumference, 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 predicts based on a specific line through the data. (See figure 2).
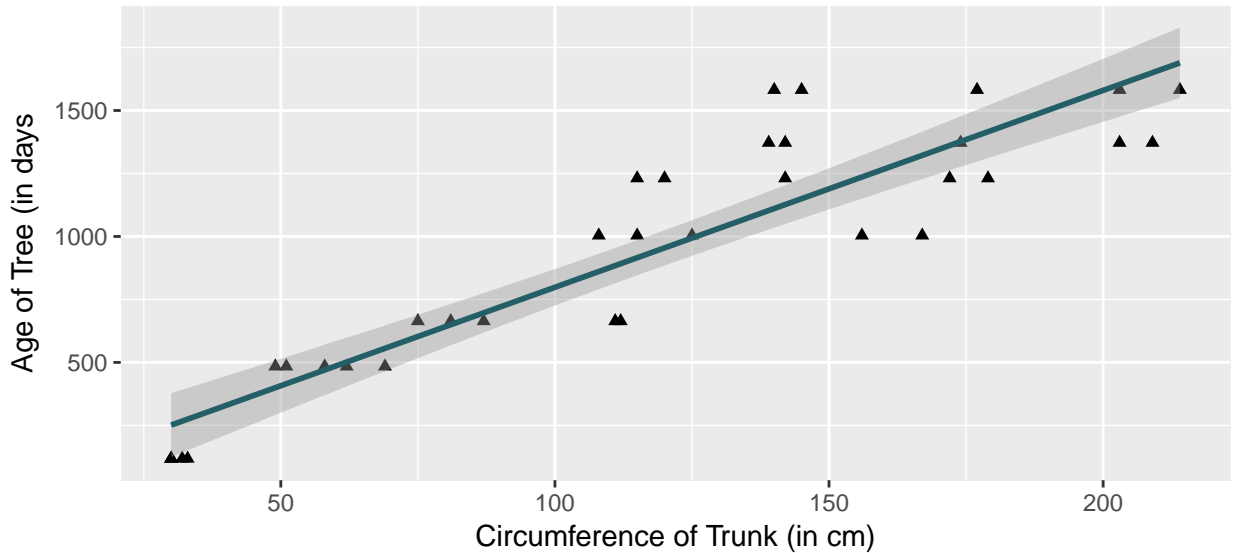


Figure 2: A linear model representing age ~ trunk circumference in orange trees. Shaded area represents a 95% confidence interval around this line.

As can be guessed from figure **??**, the linear model works better on certain data than others. Indeed, there are several assumptions that required by the linear model that may not always be present. 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, they 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 **??** and 4.
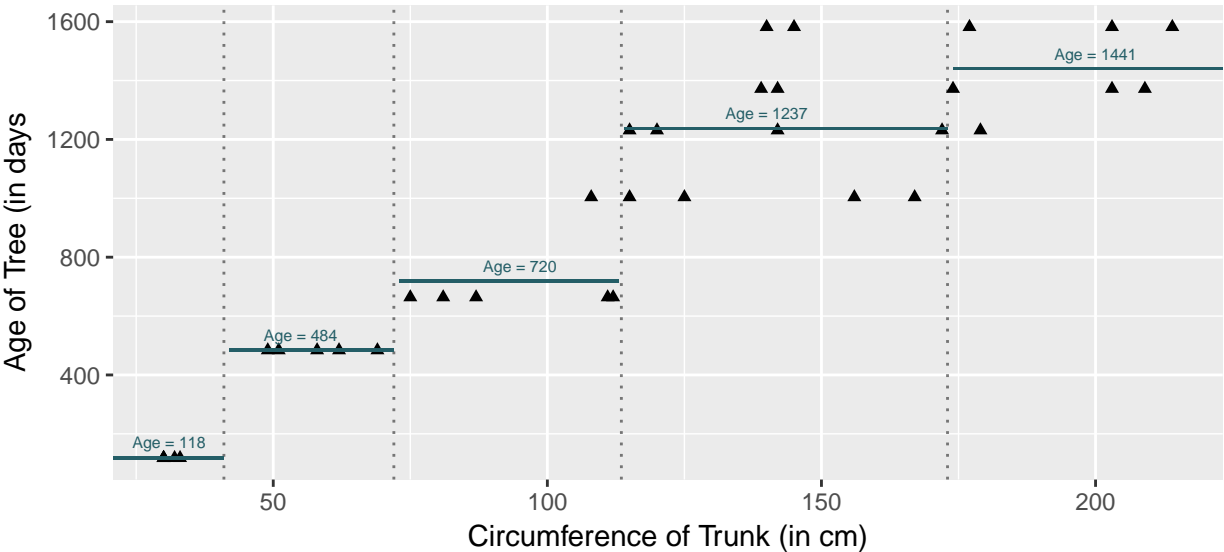


Figure 3: A tree on age ~ trunk circumference first creates partitions on the predictor, seen as vertical lines, and then predicts the value of the response within that partition, seen as text.

When using the linear model, we make predictions in the following way: given a value of X (circumference) the coresponding 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) with in the partition that X falls in to.[2]

Tree methods get their name from a common way of 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 predicitions 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 brach before landing on a prediction for age.
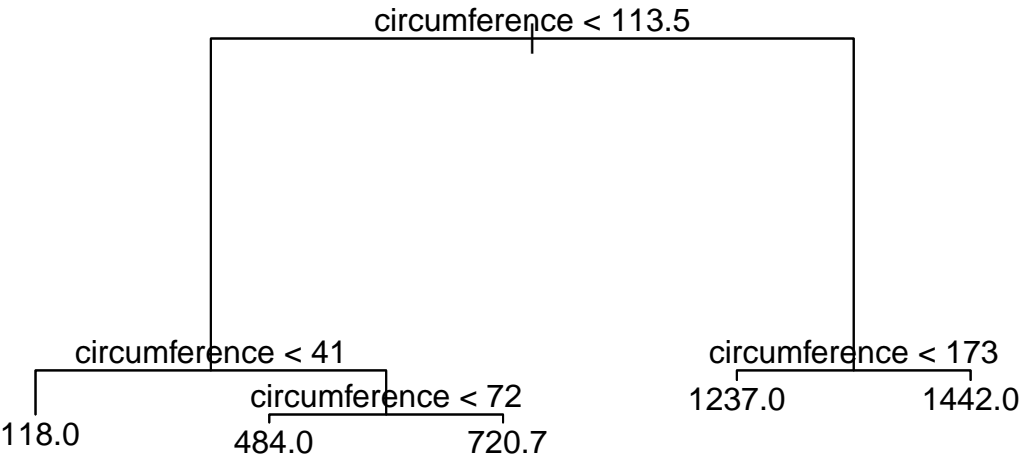


Figure 4: A tree representing age ~ trunk circumference in orange trees.

---

[2]While any curve can be approximated in a stepwise manner when the number of steps approaches $\infty$, a tree model does not converge to the linear model as the number of splits approaches $\infty$, even when the data is linear. See appendix for more.

**Trees: Background**

Decision 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 datasets Morgan and Sonquist dealt with had few participants, and a lot of data collected on each one. To add to their difficulties, there was reason to believe that there were lurking errors in the data that would be hard 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 ground work 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*, that would allow trees to be fit on various types of data. Torsten Hothorn, Kurt Hornik, 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 effect 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.

There is a limit to the predictive capabilities of a single tree as they suffer from high variance. 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. This is futher explored in Chapter 2.

**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 intepretable statistical inference. Return, for a moment, to the orange tree example from section 1.1 figure **??**.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 2: 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 gives evidence that not only is trunk circumfernce a good predictor of age, the relationship between them is the equation of the line:

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

So, 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 are not just due to chance. Inferential claims about the nature of tree age and trunk growth are possible here. A note should be made of 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 **??** 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 sample of the data, the predictions would change. See chapter 2 for more discusion on the variablility of trees. Our claims about the relation between circumference and age in young orange trees can only be descriptive as we have not taken to account the variablility in gathering them.This paper's aim is to describe a process of making inferential claims using trees and random forests that employs permuatation 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 is first 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. Then the original test statistic is tested against this distribution. If it is exceptionally rare, then there is evidence that our observation was not simply experimental variability.

## A Step Back

A random forest $R_f$ is the set of functions $T_1, ..., T_N$ where each $T_j$ is a piece-wise function from the sample space $\Omega$ into itself. In general, $\Omega$ is defined by an n x p +1 matrix where each column is a random variable.

$$when\ n = 3, p = 2$$

$$D = \begin{bmatrix} y_1 & x_{11} & x_{1p} \\ y_2 & x_{21} & x_{2p} \\ y_3 & x_{31} & x_{3p} \end{bmatrix}$$

Each tree in a random forest, $T_j \in R_f$, is generated on a subset of $\Omega$ called the training set. This training set is a bootstrapped sample of the original dataset and is noted as $B^t$. [3]:

$$B^t = \begin{bmatrix} y_1 & x_{11} & x_{1p} \\ y_1 & x_{11} & x_{1p} \\ y_3 & x_{31} & x_{3p} \end{bmatrix}$$

It is then tested on a disjoint subset of $\Omega$ called the test set, $\bar{B}^t$, where $\bar{B}^t = \Omega \backslash B^t$. The image of $T_j$ is called the predictions of $T_j$.

$$\bar{B}^t = \begin{bmatrix} y_2 & x_{21} & x_{2p} \end{bmatrix}$$

---

[3]In the bootstrapped sample, there will be repeated rows from the original data. This allows $B^t$ to be a subset of $\Omega$ without sacrificing the number of observations.

As outlined in the 1984 textbook, *Classification and Regression Trees*, Brieman, 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.

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 variablility of the whole forest. [**?**]

Predictor columns from the bootstapped sample are themselves sampled to select the columns available for a tree to split over.

$$B^t = \begin{bmatrix} y_1 & x_{11} \\ y_1 & x_{11} \\ y_3 & x_{31} \end{bmatrix}$$

## 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. At least in a single tree, descriptive claims may be simple to make, but it is much more difficult to describe the behaviour 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 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 hopes to 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 variance in model accuracy due to $X_i$. Breiman and Cutler proposed a method of permuted variable importance in their paper (cite) 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:

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

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 $\hat{*y}^t$ is the predicted outcomes after variable $X_j$ has been permuted. This value is averaged over all the trees. It's 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.

Creating a permutation-based method is certainly an attractive solution to our problem. One, it allows us to estimate the distribution of variable importance and generate a Z score under the null hypothesis that $VI_\alpha = 0$.

$$VI_\alpha(x_j) = \frac{\sum_1^n treePV^t(x_j)}{\frac{\hat{\sigma}}{\sqrt{ntree}}}$$

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.* One, 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. Secondly, the null hypothesis under Breiman and Cutler's strategy is that the variable importance $V$ 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 mislead 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 address 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.

This paper aims to provide a response to this method. The partitions are made from the random forest corresponding to the formula of $Y$ $X_1, ..., X_n$ instead of a model of $X_j$ $X_1, ..., X_n$. This ignores the common situation where the predictors are correlated enough, 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 all together.