Sean Asaro

21 April 2016

STAT 460 Project

Stone Flakes: Analysis of Prehistoric Tool Advances

INTRODUCTION

The data set *StoneFlakes* was retrieved from the Machine Learning repository at University of California, Irvine. It details the findings of anthropologists at a number of archeological sites in and around Germany and Central Europe. The researchers originally hypothesized that the stone flakes found at these archeological sites may have differed based on the technology of the society using them. After all, the stone flakes are byproducts of stone tools and the construction of other similar items from that time. From this original hypothesis, the archeologists collected a number of data points from stone flakes. These include: length-breadth index of the striking platform (LBI), relative-thickness index of the striking platform (RTI), width-depth index of the striking platform (WDI), flaking angle (FLA), frequency of platform primery (PSF), frequency of platform facetted (FSF), Dorsal surface totally worked frequency (ZDF1), and proportion of worked dorsal surface (PROZD). Along with these measurements, the data set contained a number of annotations such as the age of the site, where the site was located, and the group of Hominids most likely to be associated with the flakes based on a number of other factors. The last variable is important because it will be the response in most of the analysis. The goal of the project will be to see if the data reflects an increase of technology over time as we see early Hominids be replaced by Neanderthals and eventually Homo sapiens.

Through a series of techniques I will attempt to answer this question. The main procedures used will be clustering and classification. By the end of the project I hope to have answered these questions:

1. Can the data be clustered in a meaningful way?
2. Can classification be achieved and does it help further the analysis?
3. Is there an overall trend with stone flake characteristics and technological advances over time for Hominids?

METHODS

I will begin with clustering to attempt to answer research question 1 (RQ1). From the data I used the Euclidean distances from the original *StoneFlakes* set to create a dendrogram (Figure 1.).

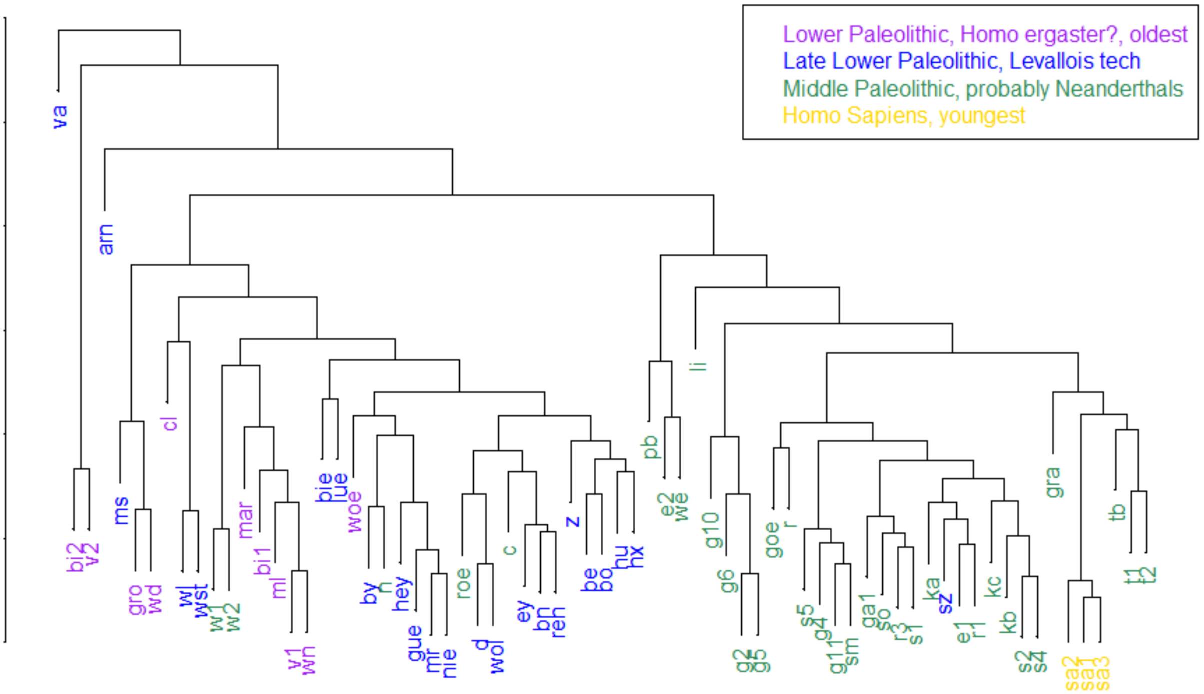


Figure 1.

You can see from the coded text labels that there is some obvious grouping just based on the distances. Almost all of the purple, Lower Paleolithic group lies on the far left. The three Homo sapiens groups lie tightly together on the right side. I also attempted to find a better cluster dendrograms using complete linkage and then Ward linkage. This gave the plots in Figures 2 and 3:

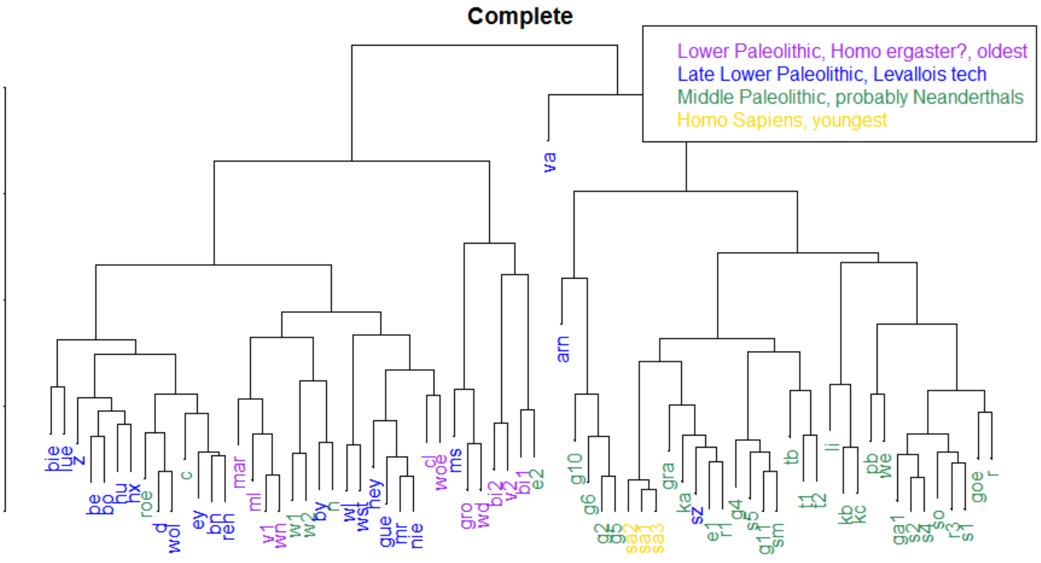


Figure 2

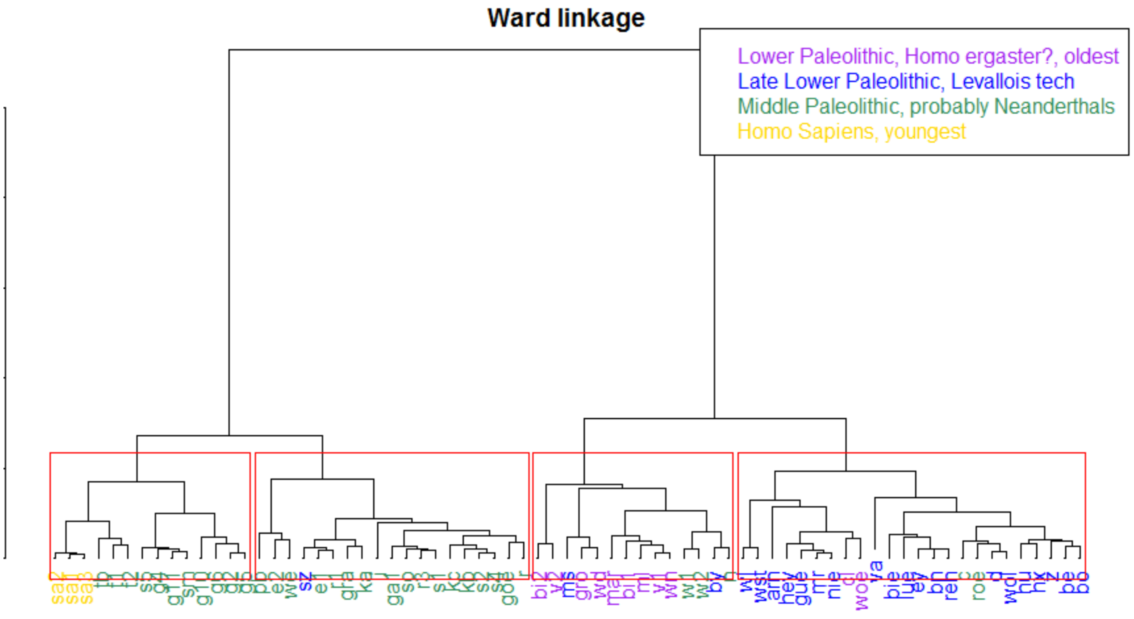


Figure 3

I believe the Ward cluster to be the best as there is a clear indication of group similarities. Most of the color-coded observation ID's are close together. However, the cluster cuts still show some error.

Following this, I also ran some Principal Component Analysis on the 8 original variables. This lead to the biplot below (Figure 4.).

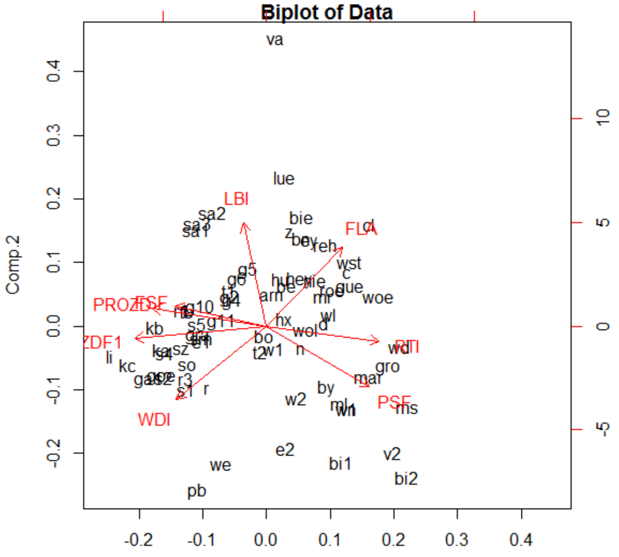


Figure 4.

The PCA narrowed the data to 2 total principle components based on the transformation and its loadings. From this I was able to plot the data in two dimensions. The loadings are detailed in the data below:

	PC1	PC2
LBI	0.08353495	0.62934222
RTI	0.40814766	0.09025955
WDI	0.32520276	0.44596791
FLA	0.27519939	0.48080250
PSF	0.37225126	0.36452757
FSF	0.33009028	0.12286919
ZDF1	0.47592915	0.07252034
PROZD	0.41340865	0.11174422

In the next section, I attempted to run classification on the data using the time periods of relative evolution as a grouping variable. From the classification tree, in Figure 3 below, one can see a fairly accurate classification structure.

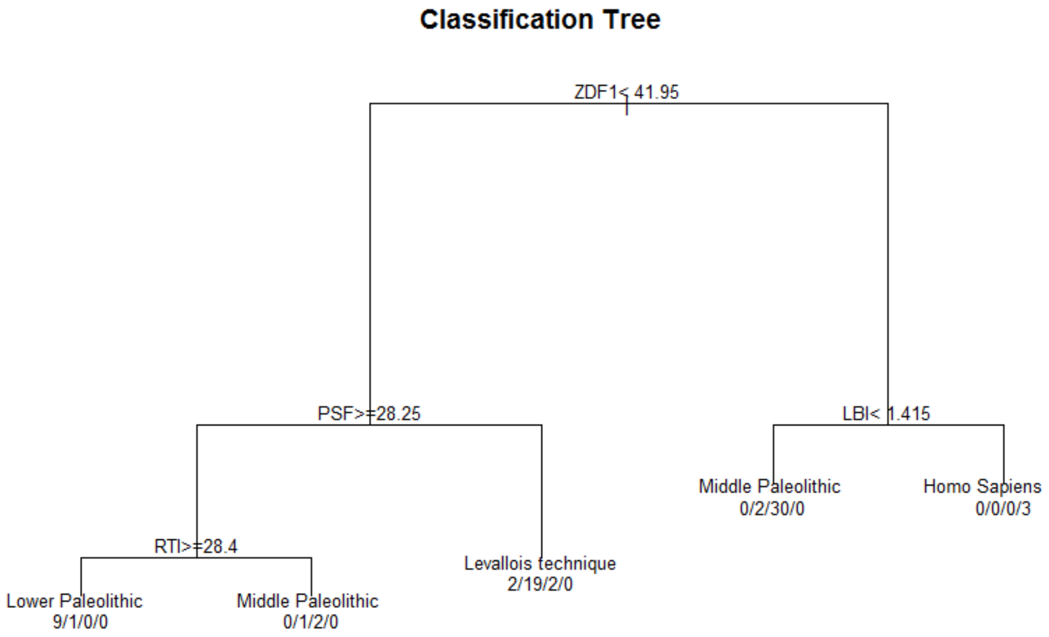


Figure 5.

Furthermore, I ran some tests to find the classification error rates for my model. Using Linear Discriminant Analysis, the re-substitution error rate came in at about 11.3%. The Leave-one-out cross validation error rate was about the same. The variables used in the tree to make these cuts may turn out to be significant in later analysis.

Finally, I added some Bagging and RandomForest models to my analysis. The Variable Importance plots for those are shown below in Figures 6 and 7:

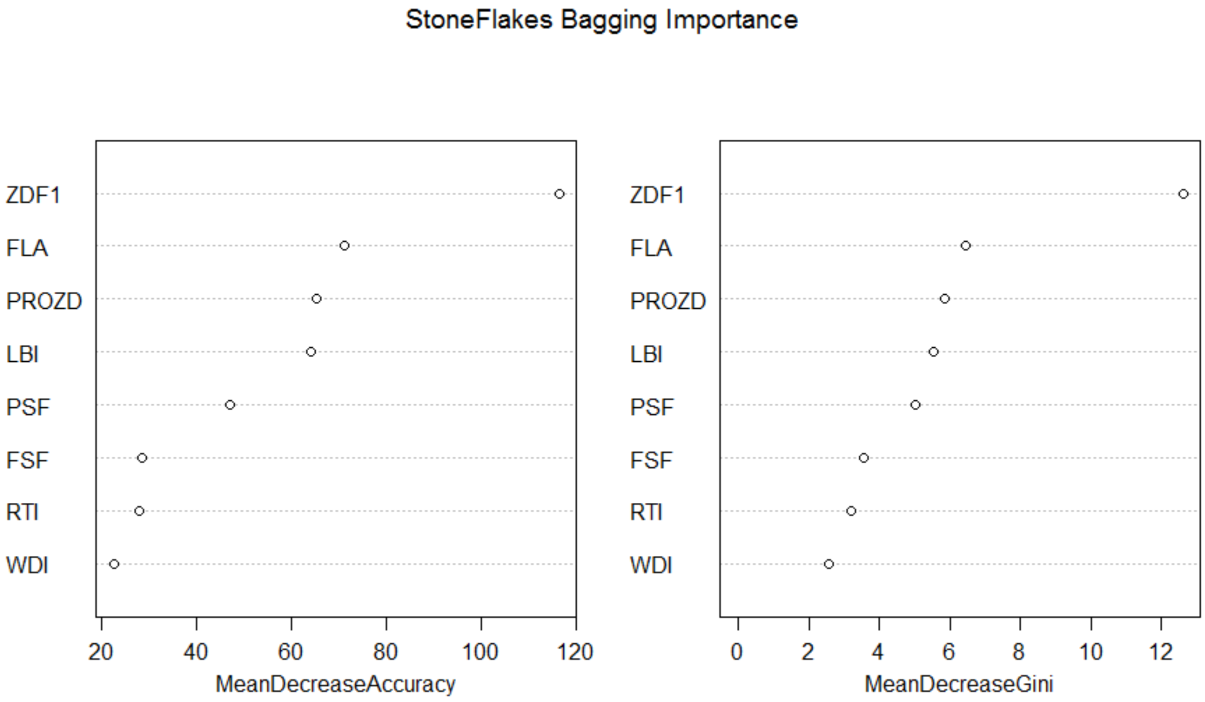
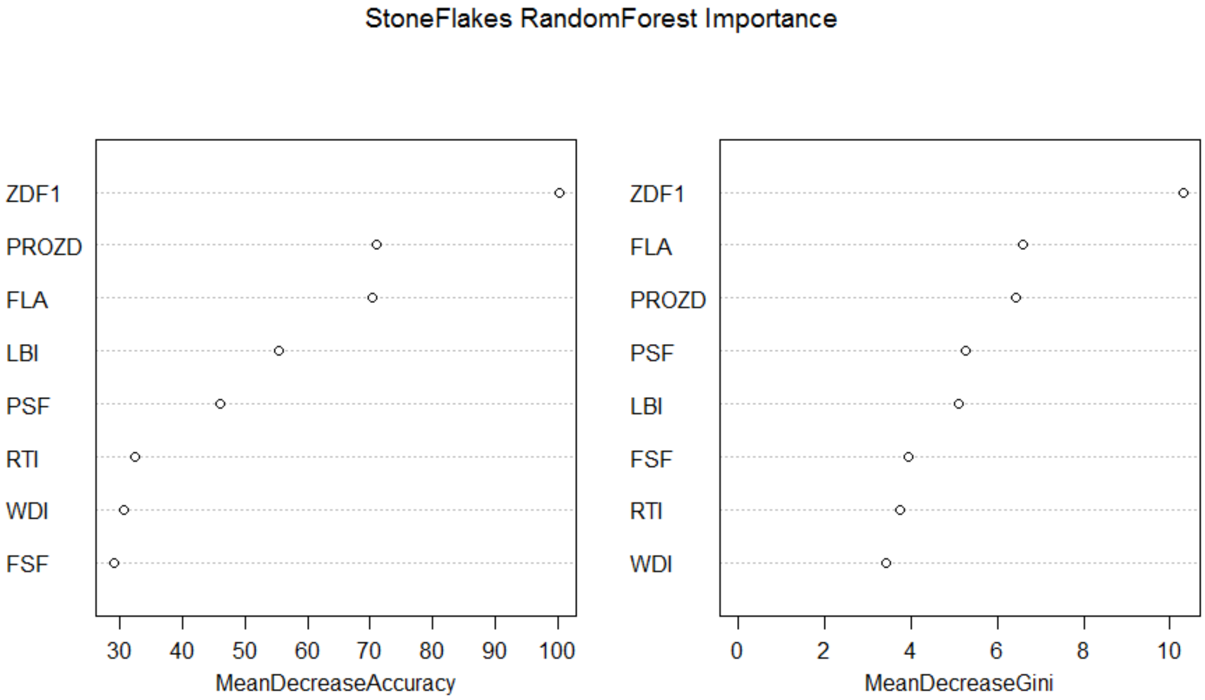

StoneFlakes Bagging Importance
ZDF1
FLA
PROZD
LBI
PSF
FSF
RTI
WDI
20
40
60
80
100
120
MeanDecreaseAccuracy
ZDF1
FLA
PROZD
LBI
PSF
FSF
RTI
WDI
0
2
4
6
8
10
12
MeanDecreaseGini

Figure 6


StoneFlakes RandomForest Importance
ZDF1
PROZD
FLA
LBI
PSF
RTI
WDI
FSF
30
40
50
60
70
80
90
100
MeanDecreaseAccuracy
ZDF1
FLA
PROZD
PSF
LBI
FSF
RTI
WDI
0
2
4
6
8
10
MeanDecreaseGini

The Bagging model showed that the most important variables were ZDF1, and then PROZD, FLA, and LBI. This model gave an error rate of 16.9%. The Random Forest model also showed the most important variable of ZDF1, followed by PROZD and FLA again. The RF model had an error rate of 19.7%.

CONCLUSION

From the clustering analysis, there are some obvious grouping variables that show up in the data. In the principal component analysis, it was apparent that some variables such as LBI, PROZD, FLA, and ZDF1 would be a strongly significant variable in grouping. In classification, the Linear Discriminant Analysis provided a fairly good classifier for Hominid groups. It is clear that there is some correlation between the observed variables and the theorized technological improvements in stone tools in Europe over the time frame. My original research question was to find a significant trend in the variables to suggest technological improvement over time. This goal was achieved. My goals of finding meaningful clusters and classification models was also successful.

REFERENCES

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Weber, Thomas. "The Lower/Middle Palaeolithic transition - is there a Lower/Middle Palaeolithic transition?" *Museo Tridentino di Scienze Naturali, Trento.* Web. 2009.