# Research Log

*Rick Venema*

*2018-09-24*

## Contents



## General introduction

In this research log, research is described that was carried out at the Hanze University of Applied sciences. This research is an attempt to create a classifier based on a forest data set. This dataset was used by Johnson et al. (2012) [1], their goal was to map different types of forests by using spectral data. The data set is available via the link https://archive.ics.uci.edu/ml/datasets/Forest+type+mapping [2].

### About the research

In the study by Johnson et al. (2012)[1], geographically weighted variables calculated for two tree species were used in addition to spectral information to classify the two different tree types in a mixed forest. Numeric spectral values were used as trainings samples for predicting spectral values at other locations using the inverse distance weighting (IDW) interpolation method. Next, the similarity between spectral values and their IDW predicted values was calculated for both tree species. This simularity is considered geographically weighted because nearer training samples have more of an impact on their calculations. This resulted in an increase in overall accuracy.

# Research question

possible to expand this to be used in different situations. Crop recognition for example, differences in one crop.

Is het mogelijk om deze techniek toe te passen in de praktijk op agrarisch niveau en kan hier mee een nauwkeurige bepaling gedaan worden.

Het doel van dit project is om een machine learning model te maken dat kan bepalen wat voor soort bomen er in een bos staan. Verder is het doel om te kijken of het model statistisch gezien sterk genoeg is om in andere toepassingen dienst te doen, bijvoorbeeld in de landbouw, dit omdat deze techniek nog in de kinderschoenen staat in de landbouw. Een model dat kan bepalen om wat voor soort het gaat, kan als opstapje dienen om een universeel model te maken dat alles kan herkennen in een afbeelding.

Can a ML algorithm be accurate for recognizing different tree types in a forest, and can it have a practical use in other fields.

Can a ML alg. give accurate results in recognizing different tree types in a forest.

# Dataset

## Dataset description

This description is the description included with the site at which the data can be found, it gives a clear description on what kind of data we are dealing with.

`This data set contains training and testing data from a remote sensing study which mapped different forest types based on their spectral characteristics at visible-to-near infrared wavelengths, using ASTER satellite imagery. The output (forest type map) can be used to identify and/or quantify the ecosystem services (e.g. carbon storage, erosion protection) provided by the forest.`

## Layout of the data

The dimensions of rawDataTesting is equal to *325* and *28*. The first value represents the amount of rows and the second value represents the number of columns. The rows represent the instances, while the columns represent the attributes. The dimensions of rawDataTraining is equal to *198* and *28*.

### The class attribute

The data has a class attribute, this class attribute can have different values which are presented in table 1

Table 1: Class Attributes

| Abbrevation | Full name |
|---|---|
| s | 'Sugi' forest |
| h | 'Hinoki' forest |
| d | 'Mixed deciduous' forest |
| o | 'Other' non-forest land |

#### The b columns

In figure 1 the log2 transformed data of the b columns. The boxplot give a difference in the different years, this is because the different columns can be divided into 3 different years. The first 3 boxplots represent the data from 26 September 2010, the next three boxplots represent the data from 19 March 2011, and the last three boxplots represent the data from 8 May 2011. The different colors of the boxplots represent the different spectral wavelengths. The green color represent their corresponding wavelength green(0.52-0.60 µm), red (0.63-0.69 µm) and near-infrared (NIR) (0.76-0.86 µm). NIR is represented as violet in the boxplots.

#### The b columns orderd by class

To get a better view of the differences between dates, the data was split into the 4 different class attributes. Each different kind of forest was put into a boxplot, which can be seen in figure 2. As can be seen, the dates differ a lot. This can be due to the different seasons at which the pictures were taken. The boxplots created from the March 2011 data, is higher than the other dates. THe data needs to be divided into the 3 groups. Each data has 3 boxplots, which each corresponding to the different spectral bands that are used in the research. These bands differ quite a bit, just like the boxplots in figure 1. This can be due to the different seasons at which the images were taken. Different seasons have different values for the different spectral bands. The spectral bands are collected by ASTRAL imagery.

#### pred_minus_obs columns

The `pred_minus_obs` columns represent the Predicted spectral values minus actual spectral values for the `s` or `h` class. These values are calculated based on spatial interpolation. For this calculation the IDW method was used, this method considers the values of nearby samples to predict the value at a given location.

# EDA

## Evenly or unevenly classes

The classes are unevenly represented. The dimensions are printed next, ordered by class type

Hinoki amount: *48*, Sugi amount: *59*, Mixed amount: *54*, Other amount: *37*

## Missing data

To check if there is missing values the dataset needs to be searched for NA values.

Is any number NA? *FALSE*

This check has been done by using the `any()` function and the `is.na()` function.

## hierarchical clustering

A hierarchical clustering was performed to get a view of the data clustered into the different groups. Each group is clustered into different groups. As figure 3 shows, the data is clustered mostly in the correct groups, however some of the data is not clustered right. This can have different reasons, especially when the Mixed class and Other class are not correctly clustered.

## PCA

A Principal Component Analysis (PCA) was performed on the data. PCA is a dimensionality reduction technique that is used in data analysis. Reducing the dimensionality of a dataset can be useful in different ways[3]. First a summary of the data was made. The three different rows per date were summed to get one column per date. This reduced the spread of the PCA.

Table 2: Summary of the PCA data

| 26 September 2010 | 19 March 2011 | 8 May 2011 |
|---|---|---|
| Min. :112.0 | Min. :194.0 | Min. :131.0 |
| 1st Qu.:143.0 | 1st Qu.:239.2 | 1st Qu.:165.0 |
| Median :164.0 | Median :258.0 | Median :180.0 |
| Mean :167.6 | Mean :260.8 | Mean :180.4 |
| 3rd Qu.:179.0 | 3rd Qu.:277.8 | 3rd Qu.:193.0 |
| Max. :461.0 | Max. :340.0 | Max. :287.0 |

Next a Prediction was executed and the results are shown in 3.

Table 3: Prediction of the log2 data

| | PC1 | PC2 | PC3 |
|---|---|---|---|
| **197** | -0.9606 | 0.06507 | -0.05448 |
| **198** | 0.2011 | -0.1424 | -0.1697 |

# References


[1] Johnson, B., Tateishi, R., Xie, Z., 2012. *Using geographically-weighted variables for image classification. Remote Sensing Letters*, 3 (6), 491-499.

[2] *Forest type mapping Data Set* Retrieved September 11, 2018, from https://archive.ics.uci.edu/ml/datasets/Forest+type+mapping

[3] *Introduction to Principal Component Analysis (PCA)* Viewed September 24 2018, from https://tgmstat.wordpress.com/2013/11/21/introduction-to-principal-component-analysis-pca/

## Attachments

### Figure 1

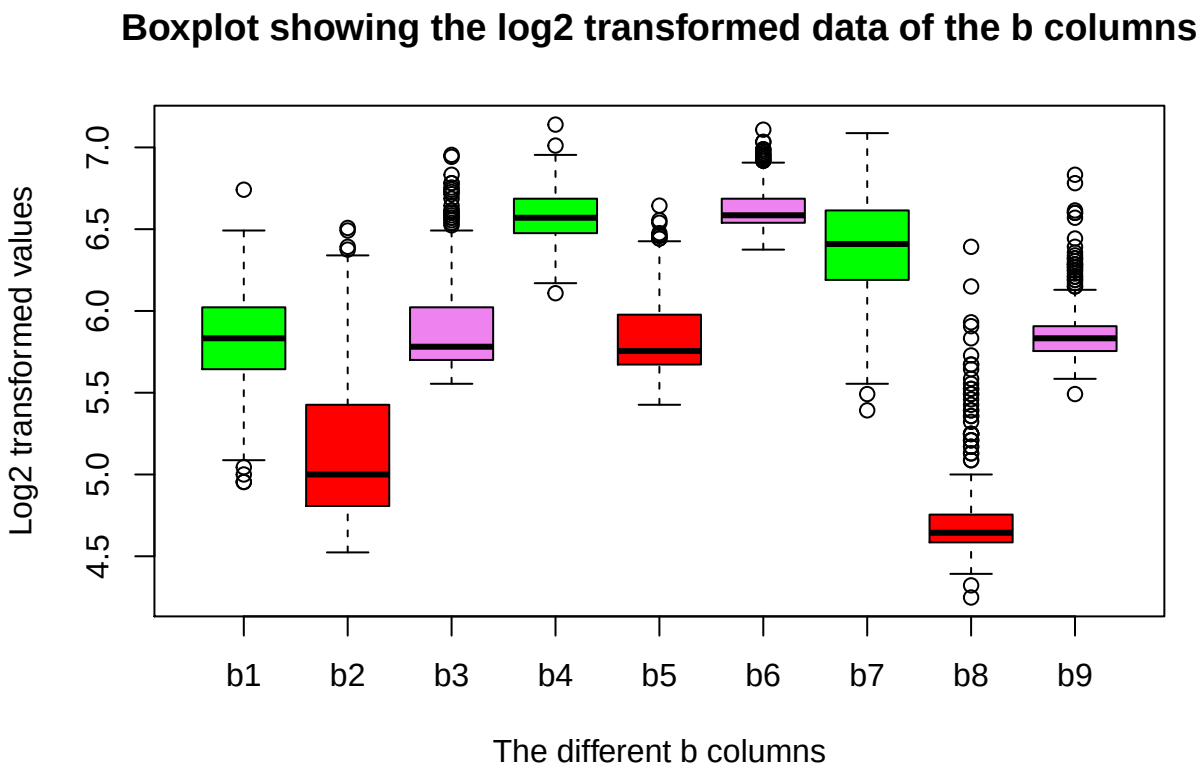


Figure 1: The log2 transformed values of the b columns.

**Figure 2**

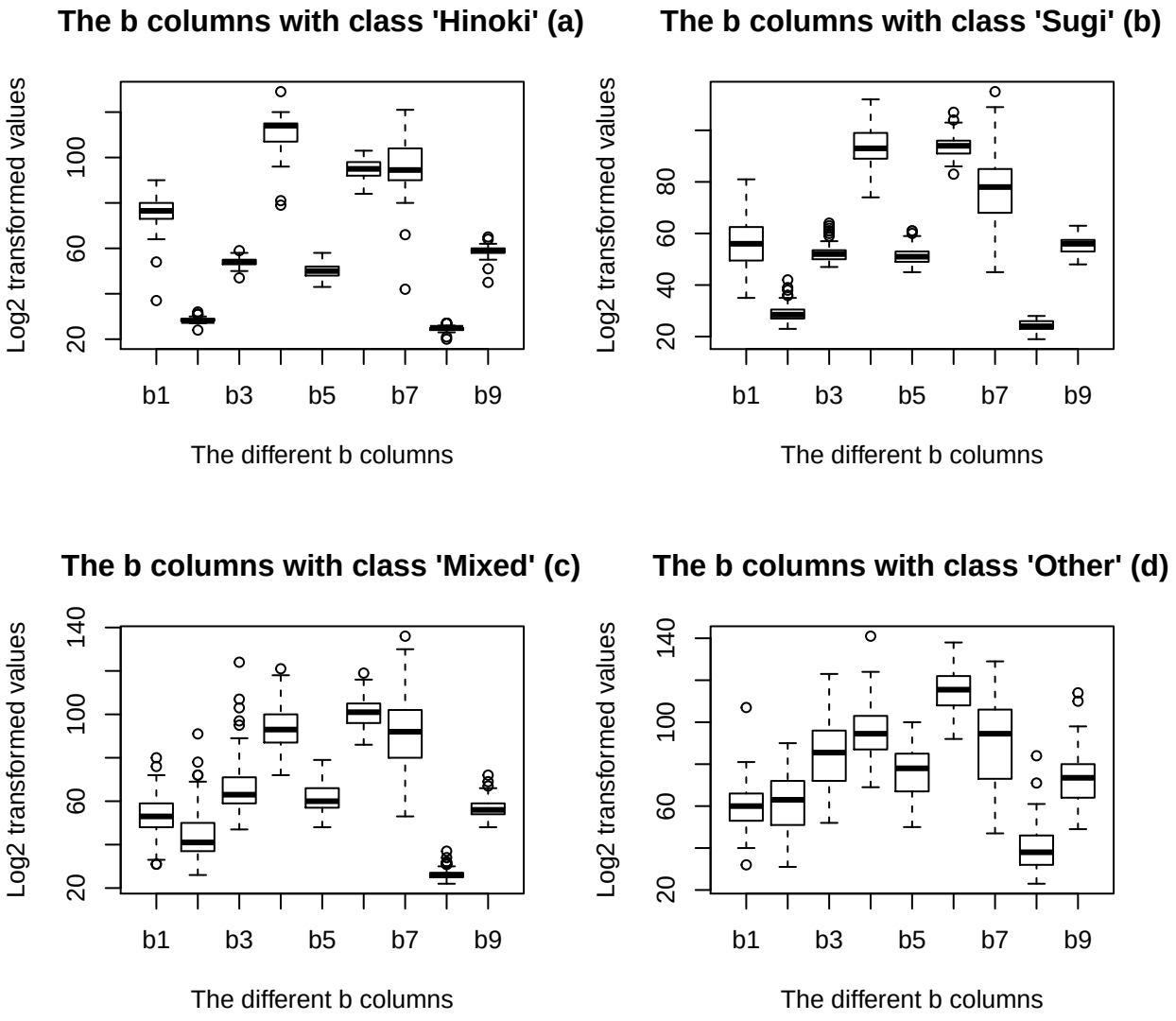


Figure 2: B columns divided by class attribute

**Figure 3**

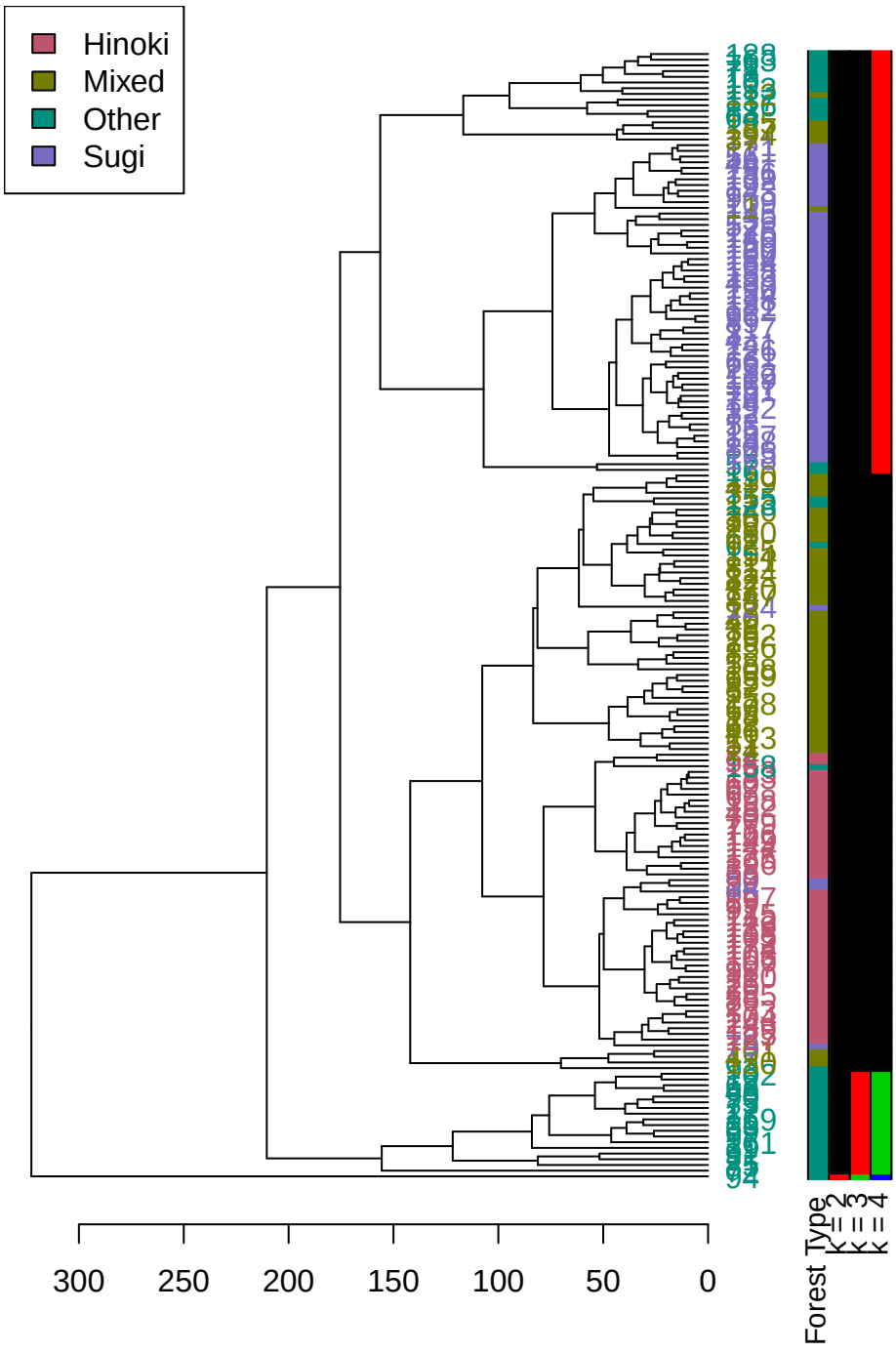


Figure 3: Hierarchical clustered data of the training Data

**Figure 4**

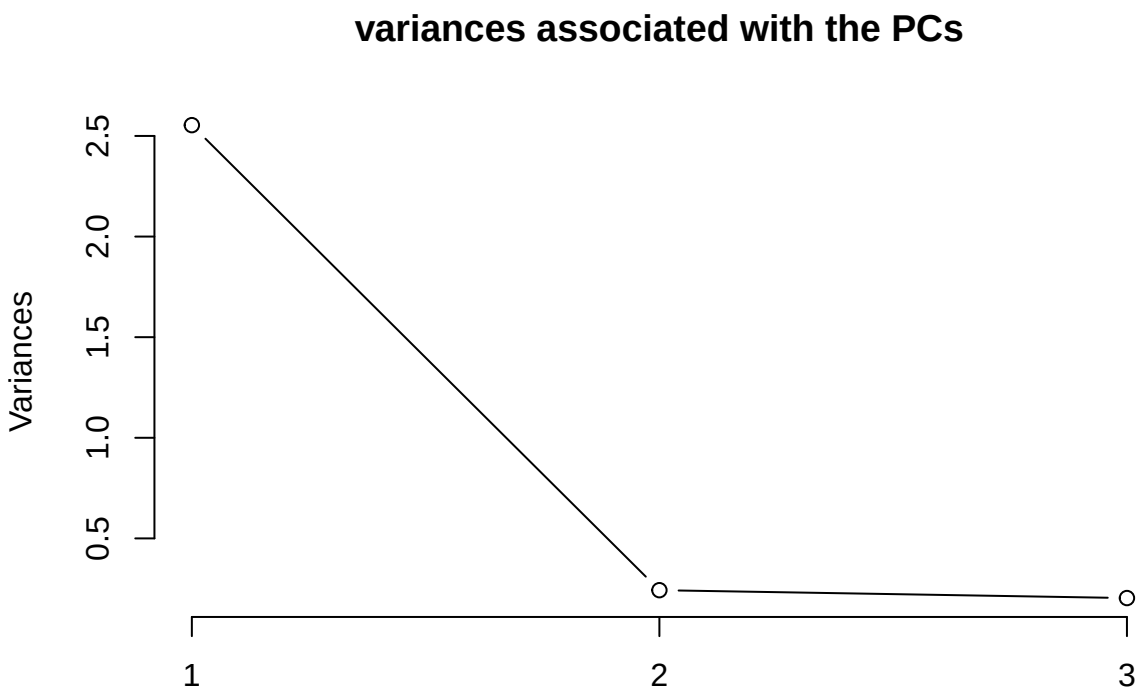


Figure 4:

**Figure 5**

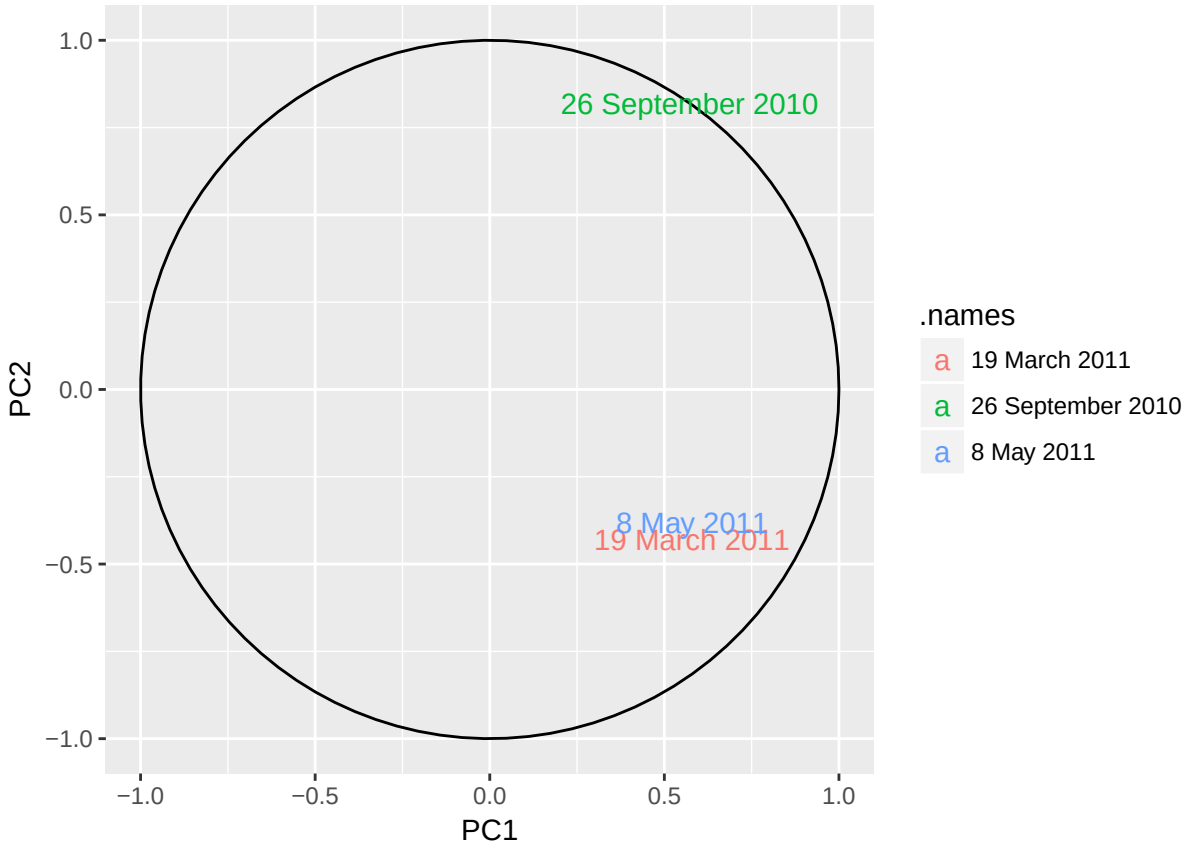


Figure 5: