Research Log

Rick Venema

2018-09-27

Contents

1 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].

1.1 License

1.2 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.

2 Research question

Based on the research chosen, a research question was created to further improve a model that can classify different forest types in a mixed forest. The question that was eventually created needed to include ML algorithms. The first question that was created was validated by classmates.

```
Can a Machine learning algorithm give accurate results in recognizing different tree
types in a forest.
```

This question had different comments, that were taken into account in improving the research question.

```
What difference can a machine learning algorithm produce when recognizing different
tree types in a forest, taking geographical weighted variables into account?
```

3 Dataset

3.1 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.
```

3.2 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*.

3.2.1 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

3.2.2 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 ?m), red (0.63-0.69 ?m) and near-infrared (NIR) (0.76-0.86 ?m). NIR is represented as violet in the boxplots.

3.2.3 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.

3.2.4 Plots

3.2.4.1 Figure 1

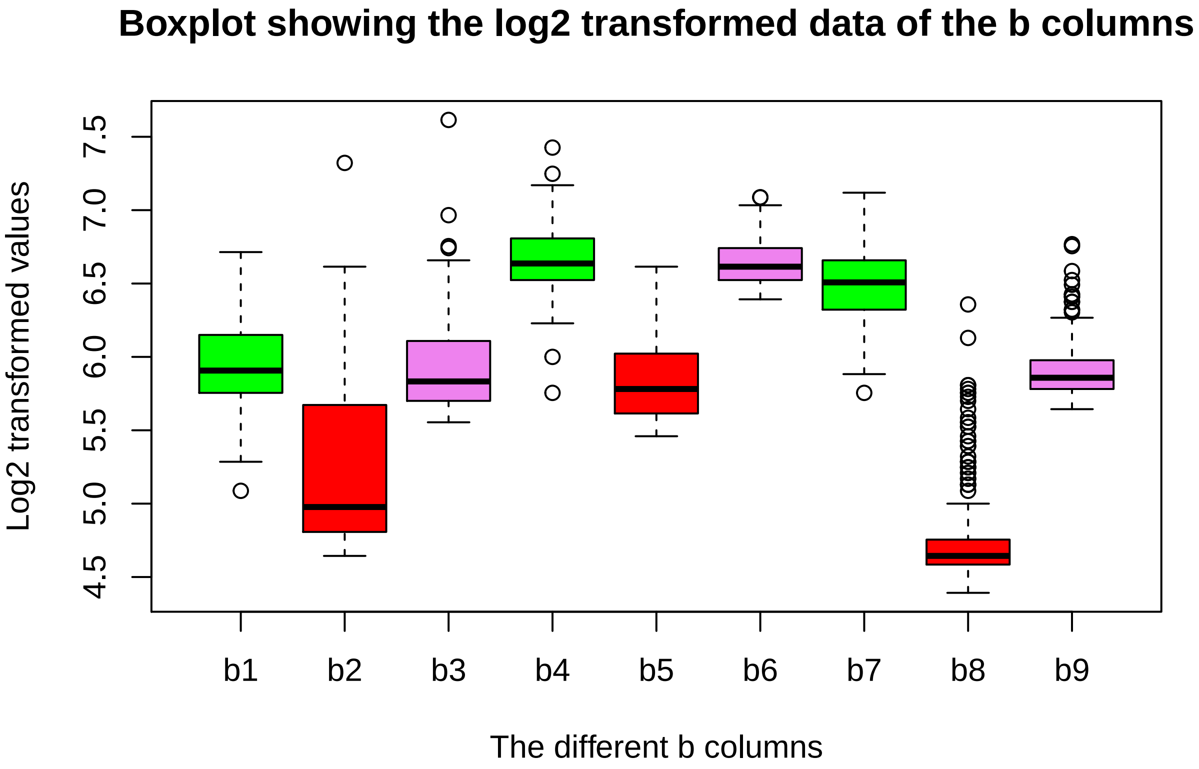

Figure 1: The log2 transformed values of the b columns.

3.2.4.2 Figure 2

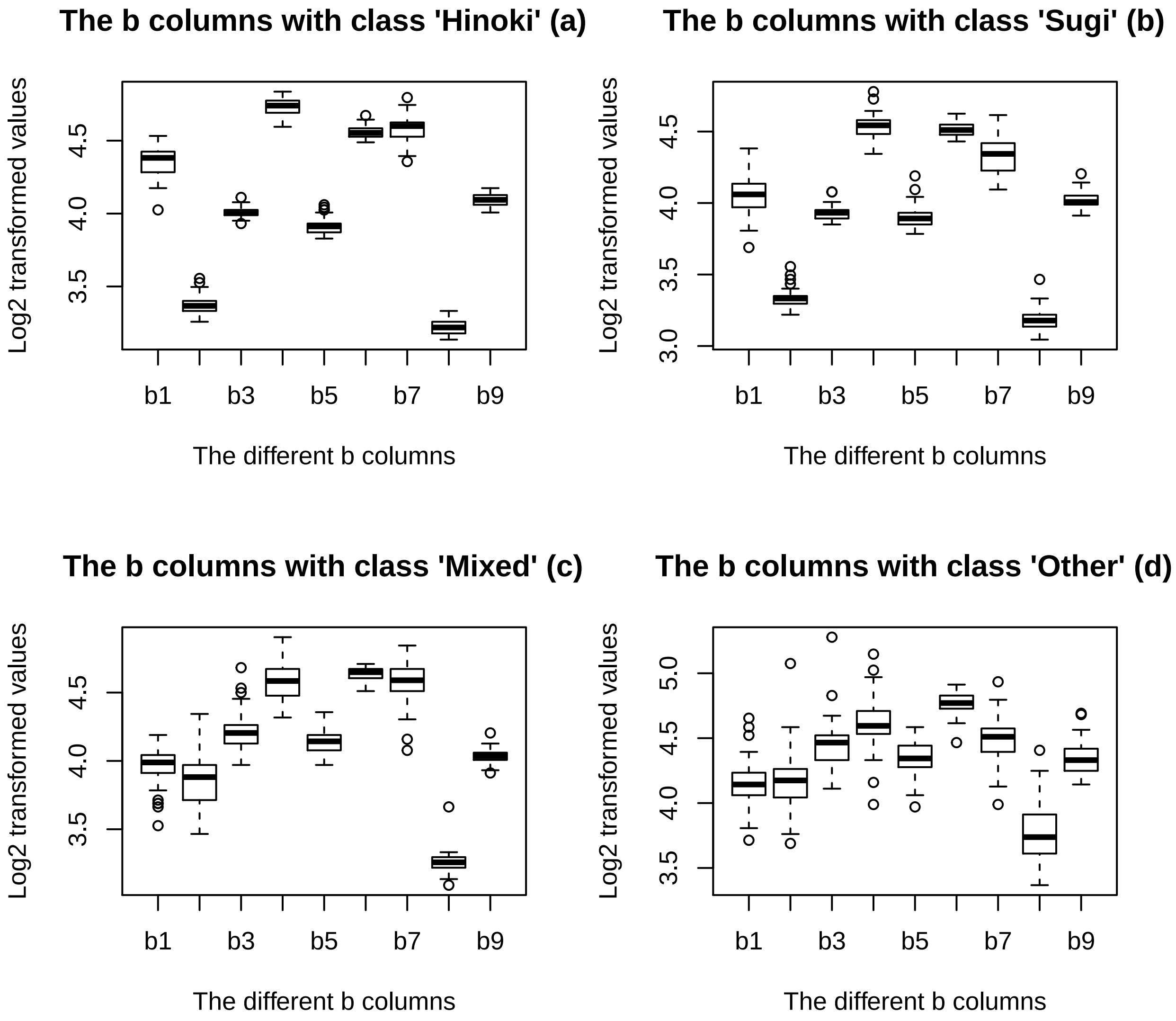

Figure 2: B columns divided by class attribute

3.2.5 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. The added 18 columns are used to determine if geographically weighted variables influences the accuracy.

4 EDA

4.1 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*

4.2 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.

4.3 Correlation in the data

The data has some correlation, the different columns respresent the different types of wavelength per date. The columns can be summed per 3 columns, the b1-b9 represent 3 dates evenly distributed. Meaning that 9 columns are a representation of 3 different dates. The next 9 columns are IDW interpolated dates in the same distribution of the first 9 b columns.

4.4 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.

4.5 Plots

4.5.1 Figure 3

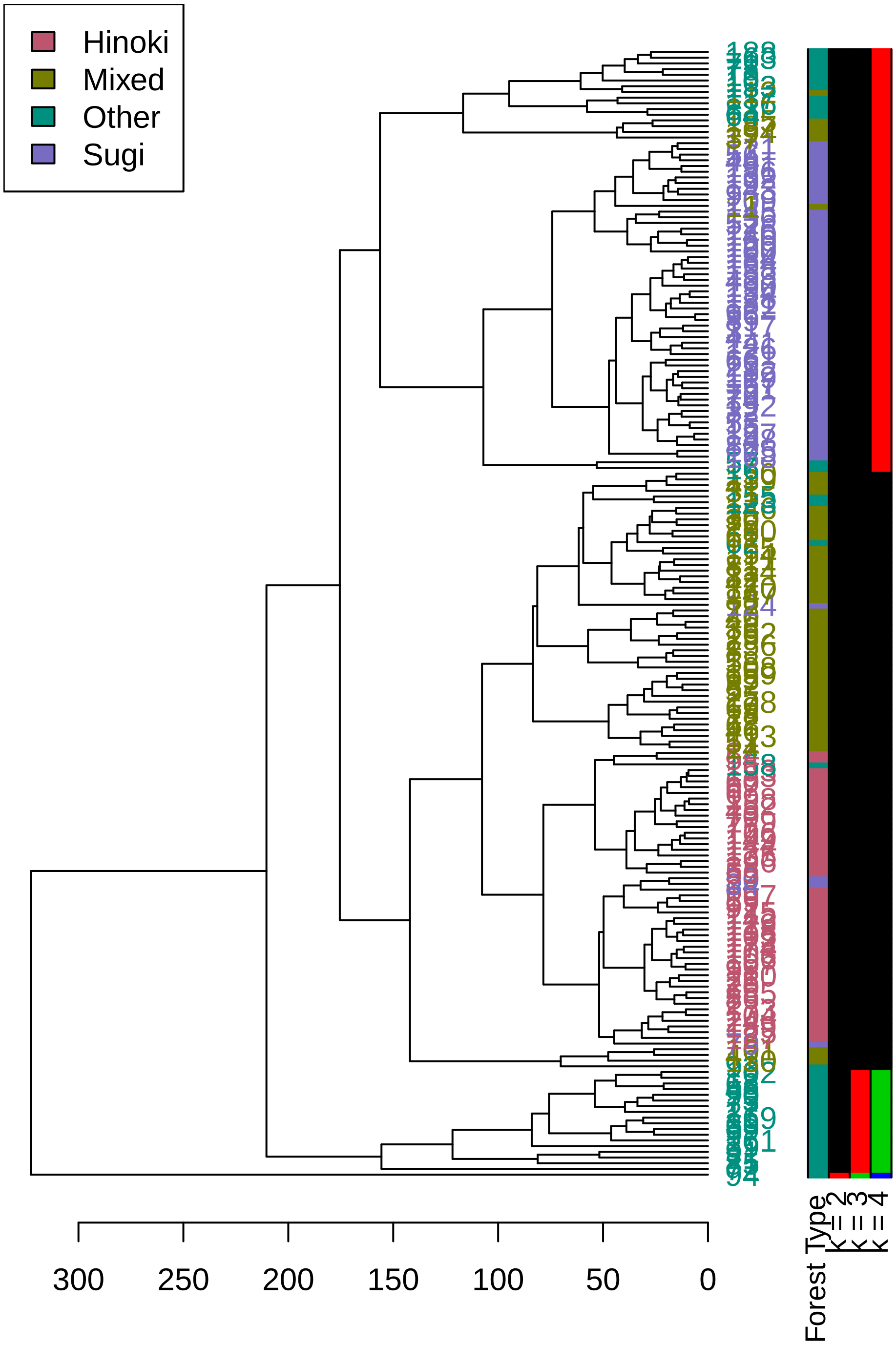

Figure 3: Hierarchical clustered data of the training Data

4.6 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

4.7 Plots

4.7.1 Figure 4

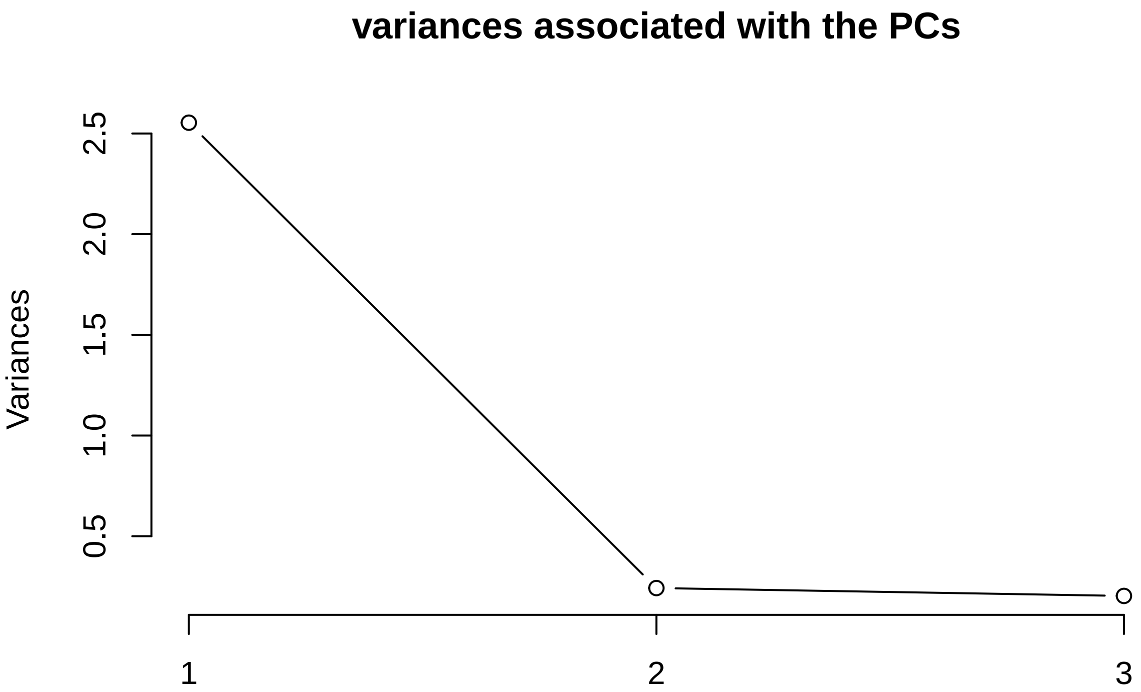

Figure 4:

4.7.2 Figure 5

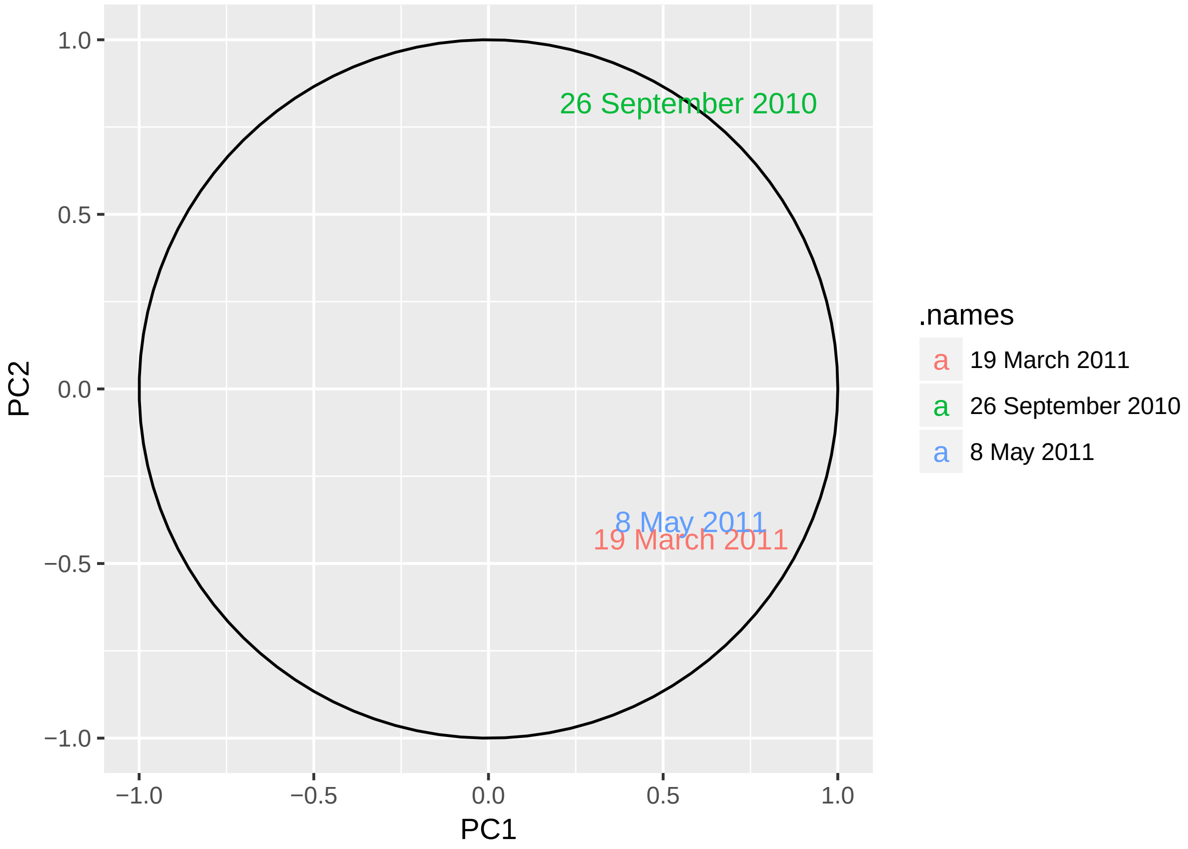

Figure 5:

5 Creating datasets

5.1 In the research

In the research the complete data set was used for training, the non interpolated data was used for reference to determine if geographical weighted values have an influence in the accuracy in classification.

5.2 Own Dataset

To create an own dataset, a couple different things need to be taken into account. For example; how do I need to split my data to create a dataset that has the geographical weighted variables and a dataset that only has the spectral values of the image. This research checks the difference in geographically weighted variables and normal spectral data, compared to onl spectral data. Meaning that 2 datasets needs to be created: the total amount of data, and only the spectral data.

```
bColumnsToFile <- data.frame(rawDataTraining[2:10], rawDataTraining$class)
allDataToFile <- data.frame(rawDataTraining[2:28], rawDataTraining$class)
```

```
write.arff(bColumnsToFile, file="../data/bColumnsOnlyData.arff")
write.arff(rawDataTraining, file="../data/CompleteTrainingsData.arff")
```

5.3 Quality metrics

Determining of quality metrics for the algorithm is an important part of assessing the ML algorithm. Each application in the real world has different requirements. To determine the quality metrics of a ML algorithm, the most important requirements need to be selected. In my case, the most important metric of my ML algorithm is speed. It is quite important that the algorithm can determine forest types on the fly. When a drone/airplane is flying over a forest, the algorthm needs to determine quickly what type of forest it is. Accuracy is not the most important part of the algorithm, it doesn't need to classify each part of the forest exactly right.

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/

[4] Russell, G., Congaltol., 1988 *A Comparison of Sampling Schemes Used in Generating Error Matrices for Assessing the Accuracy of Maps Generated from Remotely Sensed Data*