Eberswalde University for Sustainable Development (HNEE)

Faculty of Forest and Environment

Nguyen Duc Viet

Project report

on the course of – Environmental Data Analysis

Lecturer: Prof. Dr. Luis Miranda

Eberswalde, 07.2022

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

DSM	Digital Surface Model
GNSS	Global Navigation Satellite Systems
IMU	Inertial Measurement Unit
LiDAR	Light Detection and Ranging
Max	Maximum
Min	Minimum
OOB	Out-of-bag error
PCA	Principal Component Analysis
SD	Standard deviation

TABLE OF CONTENTS

1. INTRODUCTION

Most environmental data are highly complex and uncertain. Environmental data analysis contains cutting-edge methods and tools which is suitable for the needs of environmental sciences and associated fields, including descriptive and explorative statistic, machine learning techniques for clustering, classification, and regression. The past few decades have witnessed a remarkable increase in interest in these methods for environmental monitoring, modeling, and decision-making (Zhang, 2017)c.

Light Detection and Ranging (LiDAR) is an active remote sensing technology that uses a sensor to produce laser pulses and uses ultra-accurate clocks to measure the return time for each beam as it travels between the sensor and targets. The location of each laser return is achieved by precise kinematic positioning using Global Navigation Satellite Systems (GNSS) and orientation characteristics received from an Inertial Measurement Unit (IMU). The increasing use of LiDAR technology has revolutionized the acquisition of precise three-dimensional and non-destructive forest structure measurements.

Based on data analysis techniques, this study aims to explore the height data from airborne LiDAR scanning and get some understanding of the studied forest environment. To do that, a dataset of 27 height metrics for 20 LiDAR scanning tiles was used. Principal Component Analysis (PCA) was performed over the 27 metrics to explore the latent dimensionality of the height dataset and select the most important principal components (PCs) carrying the most variance for further clustering and classification. The focus is on searching for any existing patterns of clustered data which could suggest more about groups of tiles that are similar in terms of height, by using unsupervised K-means clustering and supervised decision tree, random forest algorithm.

1.1. Objectives

This study conducted some specific tasks as follows:

1. Calculating 27 statistical metrics for the height data over 20 tiles
2. Performing Principal Component Analysis over the 27 metrics
3. Selecting the most useful PCs covering the most variance
4. Performing K-means clustering with selected PCs
5. Implementing decision tree classifier with selected PCs
6. Performing random forest classification with selected PCs

2. MATERIAL AND METHOD

2.1. Study area

The study area covers an area of 20 km^2 located in the Free State of Thuringia, central Germany (Fig. 1). The state covers 16,171 km^2 with a population of about 2.1 million. Due to its extensive, dense forest, it has been referred to as "the green heart of Germany" since the 19th century (Wikipedia, 2022). Thuringia's original natural vegetation was a forest with beech as the main species. A blend of beech and spruce would be typical in the uplands. However, most of the forests are spruce and pine-planted, while most of the plains have been cleared and are being used for intensive agriculture. Since 1990, Thuringia's forests have been maintained to create tougher, more natural vegetation that is resistant to pests and disease (Wikipedia, 2022). The average daily high temperature of Thuringia, one of Germany's coldest regions, is only 12^{o}C. The weather is generally consistent with that of Central Europe (Worlddata.info, 2022).

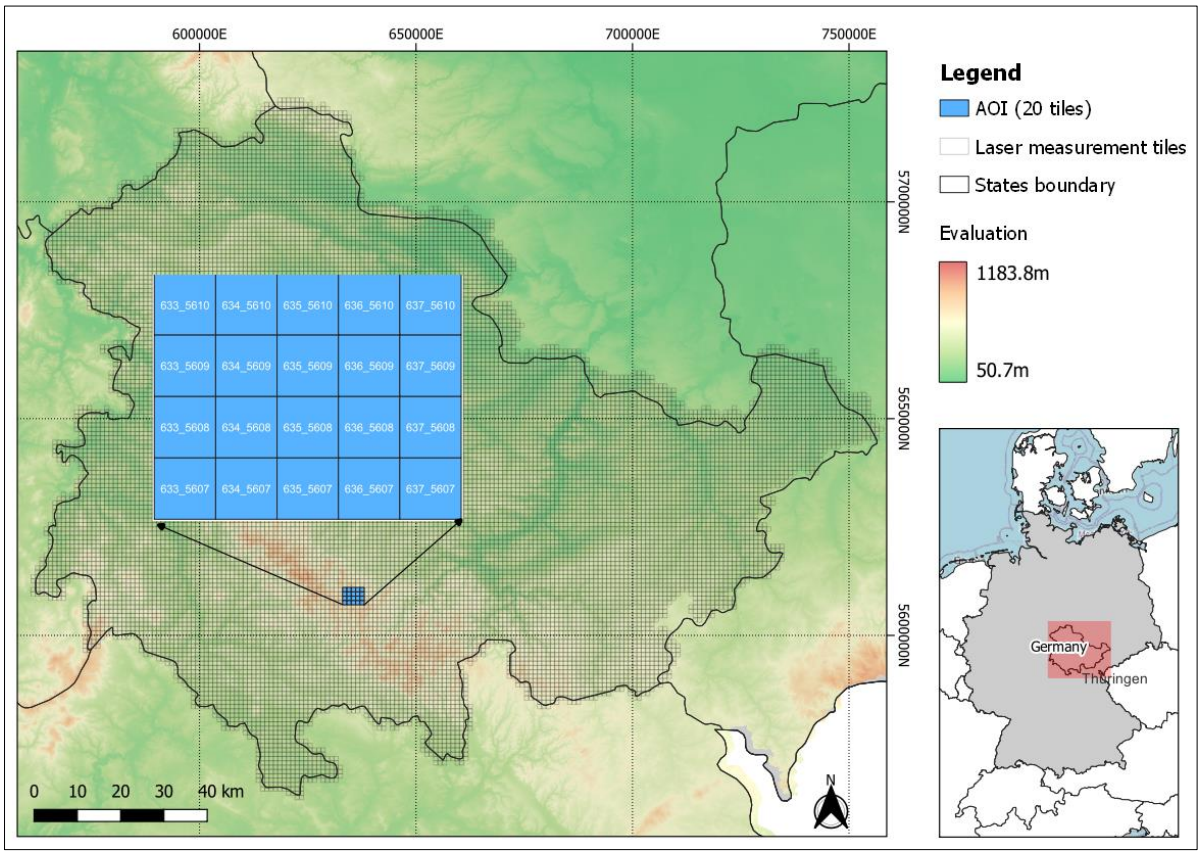

Figure 1. *Study area within Thuringia and its location in Germany. The grid shows LiDAR scanning tiles over the state. The 20 LiDAR tiles of interest were marked in blue with tile ID (section 2.2). Digital evaluation model used as background from Copernicus Land Monitoring Service.*

2.2. LiDAR data

The government of Thuringia, Geoportal of the Land Thuringia[1] , provides Digital Surface Model (DSM) based on 3D measurement data from airborne laser scanners for the entire state. The scanning data was divided into regular non-redundant tiles with an area of 1x1 km per tile. Data used in this study contains 20 tiles (20km^2) (Figure 1). The data including coordinates and height information was provided in XYZ format. There are 4 tiles were collected from 2015 and the rest from 2019 (Appendix I).

2.3. Analysis

This section describes the analysis procedures used in the study. The flow diagram outlines the methodology can be found in Figure 2. All calculations and analyses were conducted using the *R* environment with R studio 2022.07 and R version 4.2.1. Mapping was created by *QGIS 3.24.*

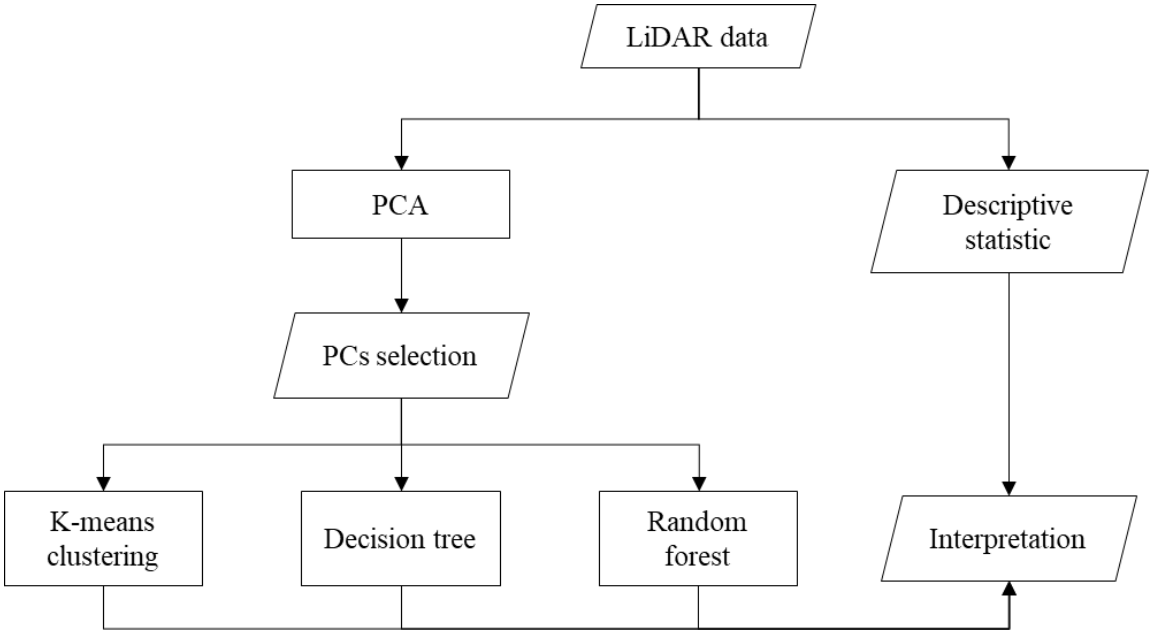

Figure 2. *Flowchart describing the process of analyzing LiDAR data*

[1] https://www.geoportal-th.de/de-de/Downloadbereiche/Download-Offene-Geodaten-Th%C3%BCringen/Download-H%C3%B6hendaten

3. RESULTS AND DISCUSSION

3.1. Descriptive statistic

Table 1 shows 27 statistic metrics calculated from height data of the 20 laser scanning tiles. Besides, the boxplot gives the first impression of differences in height data among 20 tiles (Fig. 3).

Table 1. *Statistic metrics of the height from 20 laser scanning tiles*

Tile_id	633_ 5607	633_ 5608	633_ 5609	633_ 5610	634_ 5607	634_ 5608	634_ 5609	634_ 5610	635_ 5607	635_ 5608	635_ 5609	635_ 5610	636_ 5607	636_ 5608	636_ 5609	636_ 5610	637_ 5607	637_ 5608	637_ 5609	637_ 5610
Mean	790.7	808.1	796.7	715.9	818.4	805.1	755.3	718.7	733.0	779.3	770.5	754.5	690.3	705.4	714.1	706.2	688.8	667.4	671.8	740.3
SD	29.54	19.80	34.24	39.87	24.89	19.32	44.66	60.37	35.49	24.49	36.60	34.74	21.59	31.20	29.27	36.80	27.63	38.19	49.34	31.50
Kurtosis	3.75	4.10	3.12	3.00	3.23	3.02	2.39	1.82	2.73	2.58	2.19	2.65	2.40	2.59	2.69	2.16	2.69	2.25	2.23	2.72
Skewness	-1.05	-0.86	-0.88	-0.09	-0.75	-0.43	-0.58	0.01	0.55	-0.40	-0.21	-0.32	-0.26	0.29	0.17	-0.04	0.14	0.09	-0.11	-0.61
Max	847.9	853.7	857.5	831.5	865.7	856.0	836.3	827.1	831.2	836.7	843.7	827.2	747.7	797.3	804.6	793.7	768.1	771.3	788.4	798.1
Min	679.4	724.6	674.3	602.4	722.8	724.0	630.6	594.1	664.7	701.1	673.5	654.6	635.7	628.5	633.6	616.7	618.4	585.9	569.2	637.8
Range	168.5	129.2	183.2	229.1	143.0	132.0	205.7	233.1	166.5	135.6	170.1	172.6	112.0	168.9	171.1	177.1	149.7	185.4	219.2	160.3
Interquartile	34.6	23.7	45.8	51.0	32.1	25.7	67.8	107.6	45.2	35.9	56.2	47.8	32.7	42.5	41.3	58.2	37.0	59.3	74.1	43.5
quantile_5th	727.9	771.1	726.8	646.1	768.1	770.3	670.7	625.7	680.0	735.4	707.9	693.3	651.1	656.1	668.1	646.8	643.9	606.1	584.1	682.1
quantile_10th	746.2	781.7	745.2	661.0	782.6	779.4	686.3	639.7	691.9	744.8	718.9	707.8	661.1	666.3	677.9	656.5	652.7	616.5	602.0	693.3
quantile_15th	759.0	788.3	757.2	673.2	792.7	785.1	699.1	649.6	699.2	750.8	727.5	717.2	665.7	674.3	684.2	664.6	660.3	624.2	614.9	702.5

quantile_20th	769.1	793.6	767.4	683.2	799.8	789.3	711.5	658.2	704.1	756.3	735.2	724.5	670.0	679.6	689.3	671.4	665.8	631.2	625.0	712.0
quantile_25th	776.6	798.0	777.0	691.1	805.0	792.8	722.5	667.2	708.5	762.3	742.6	731.2	674.5	684.0	693.0	677.7	670.4	638.0	635.6	720.0
quantile_30th	782.6	801.6	784.6	697.9	808.9	795.6	734.2	676.6	712.2	767.6	749.3	737.1	678.5	687.5	696.8	683.9	674.3	644.1	645.1	726.9
quantile_35th	787.5	804.4	791.0	703.9	812.3	798.7	744.6	686.1	715.7	772.1	756.2	742.7	682.2	691.3	700.7	689.4	677.5	649.9	653.2	732.6
quantile_40th	791.2	806.5	796.4	709.4	815.6	801.6	753.2	696.0	719.3	775.8	762.6	747.7	685.5	695.0	704.6	694.6	680.6	655.4	660.6	737.6
quantile_45th	795.0	808.5	801.1	713.5	818.6	804.3	760.4	706.3	722.9	779.1	768.2	752.1	688.9	698.4	708.2	700.1	683.8	660.8	667.5	741.6
quantile_50th	798.0	810.6	805.2	717.2	821.8	806.9	765.8	716.4	726.7	782.2	773.7	756.8	692.3	701.8	712.4	706.4	687.3	666.2	673.8	745.9
quantile_55th	800.7	812.6	808.9	721.3	824.8	809.4	770.2	726.8	730.8	785.2	778.6	761.0	695.4	705.6	716.3	711.7	690.8	671.8	680.3	750.0
quantile_60th	803.5	814.5	812.3	725.9	827.8	811.6	775.0	737.2	735.2	787.6	783.1	765.3	698.2	709.6	720.3	718.1	694.6	677.8	686.8	753.5
quantile_65th	805.9	816.8	815.9	731.3	830.7	813.4	779.9	748.0	740.5	790.8	787.8	769.1	700.9	714.2	724.8	724.9	698.9	684.4	693.2	756.9
quantile_70th	808.2	819.2	819.3	736.8	833.8	815.8	785.2	760.5	746.5	794.4	793.1	773.2	703.9	720.0	729.7	730.7	703.2	690.9	700.6	760.3
quantile_75th	811.2	821.7	822.8	742.1	837.1	818.4	790.3	774.9	753.7	798.1	798.8	779.0	707.2	726.5	734.3	735.9	707.4	697.3	709.7	763.6
quantile_80th	814.5	824.3	826.3	747.4	840.7	821.5	795.5	786.1	762.2	801.8	805.2	786.8	710.3	733.2	739.4	741.0	712.0	703.2	718.6	767.8
quantile_85th	817.7	827.4	829.9	753.6	844.4	825.2	800.8	795.2	774.1	805.4	811.9	794.0	713.4	741.1	746.0	746.8	717.8	709.4	727.3	773.3
quantile_90th	821.6	831.0	834.1	763.4	848.1	829.7	806.4	803.3	789.3	809.3	818.5	800.9	716.7	751.0	753.9	755.0	726.8	716.9	737.0	778.5
quantile_95th	827.1	836.0	839.1	782.8	852.1	834.9	815.2	809.8	802.9	814.9	827.7	808.2	722.0	762.2	764.6	764.7	738.0	731.9	749.7	783.7

Abbreviation: SD – standard deviation, Max – maximum, Min - Minimum

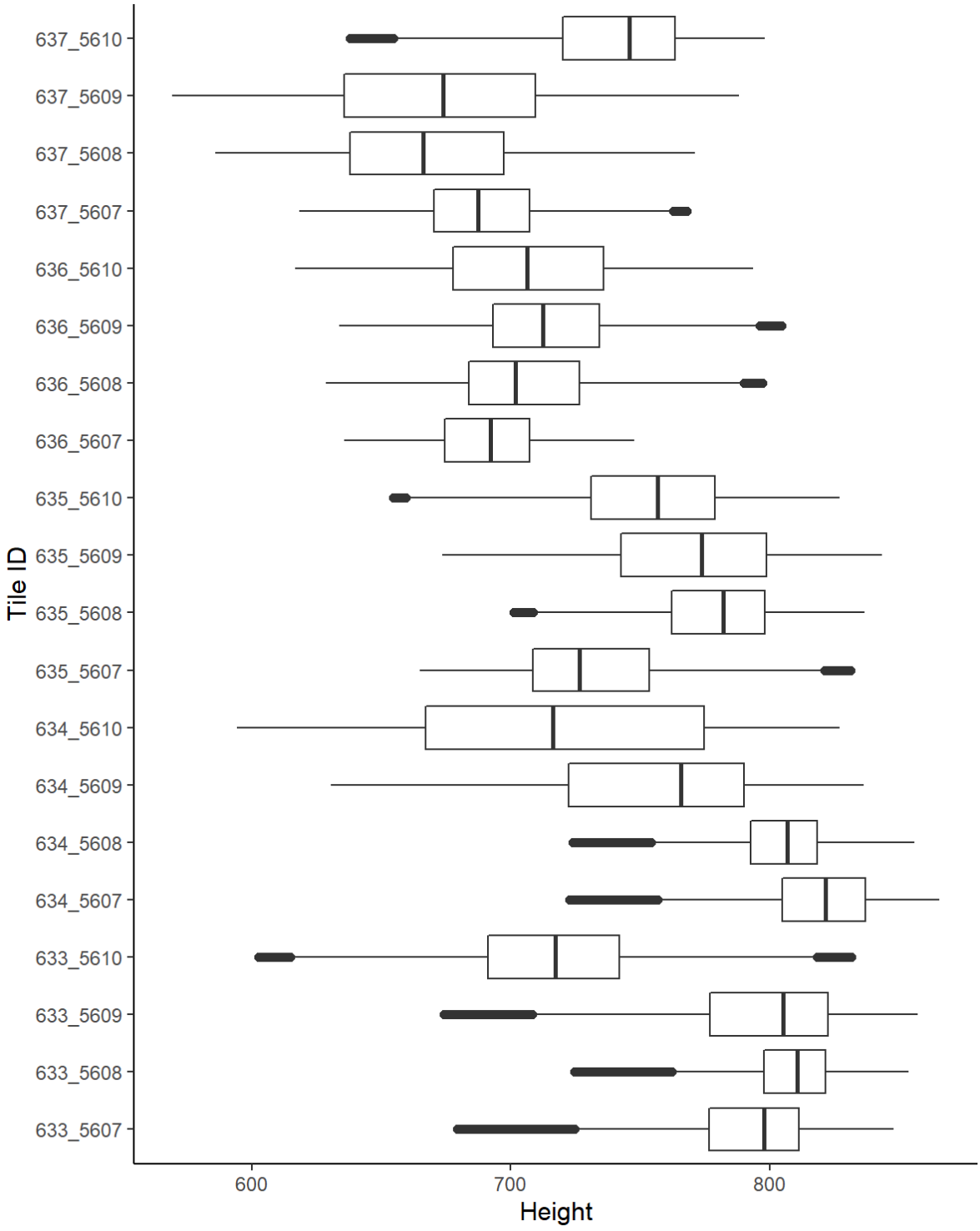

Figure 3. *Boxplots of height data from 20 laser scanning tiles*

3.2. Principal Component Analysis

PCA was performed over the 27 metrics (Table 1) to explore the latent dimensionality of the dataset. From that, the essential parts of the data that have more variation can be identified and selected for further clustering and classification. In other words, the variable that

is better in differentiating the data into groups can be determined by PCA. Scree plot using the square of standard deviation to show how much variation each PC accounts for, PCs appeared in order of the amount of variation they cover (Fig. 4).

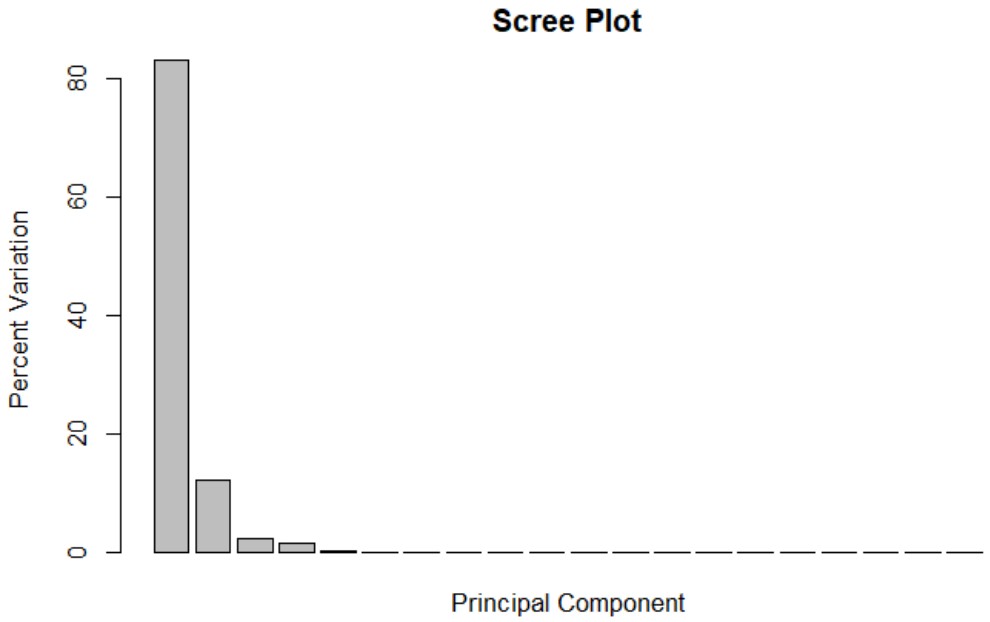

Figure 4. *Scree plot showing amount of variation each PC capture*

The result shows that PC1, PC2, PC3, PC4, PC5 and PC6 account for 83.1, 12.2, 2.5, 1.7, 0.4, 0.1 percentage of variation, respectively. Meanwhile, there is no variation accounted for by PC7 to PC20.

Moreover, biplots of PC1-PC2, PC1-PC2, and PC2-PC3 are given to illustrate how metrics distribute to each PC and how metrics correlate with each other (Figure 5). The directions of the vectors tell which PC they tend to contribute to. For instance, from Figure 5a, it can be seen that a group of quantile metrics contribute most for PC1, while the metrics that contribute most for PC2 are range and interquartile. Moreover, the length of the vector refers to how strong the vectors (metrics) contribute to the PCs. Besides, when two vectors are close, forming a small angle, they present a positively correlated, e.g., quantile metrics in PC1 (Fig. 5a). Conversely, when they form a large angle (close to 180 degrees), they are negatively correlated, for example, skewness and kurtosis in PC1 (Fig. 5a). While if they form a 90-degree angle, they are probably not to be correlated, e.g., max and range in PC2 (Fig. 5c).

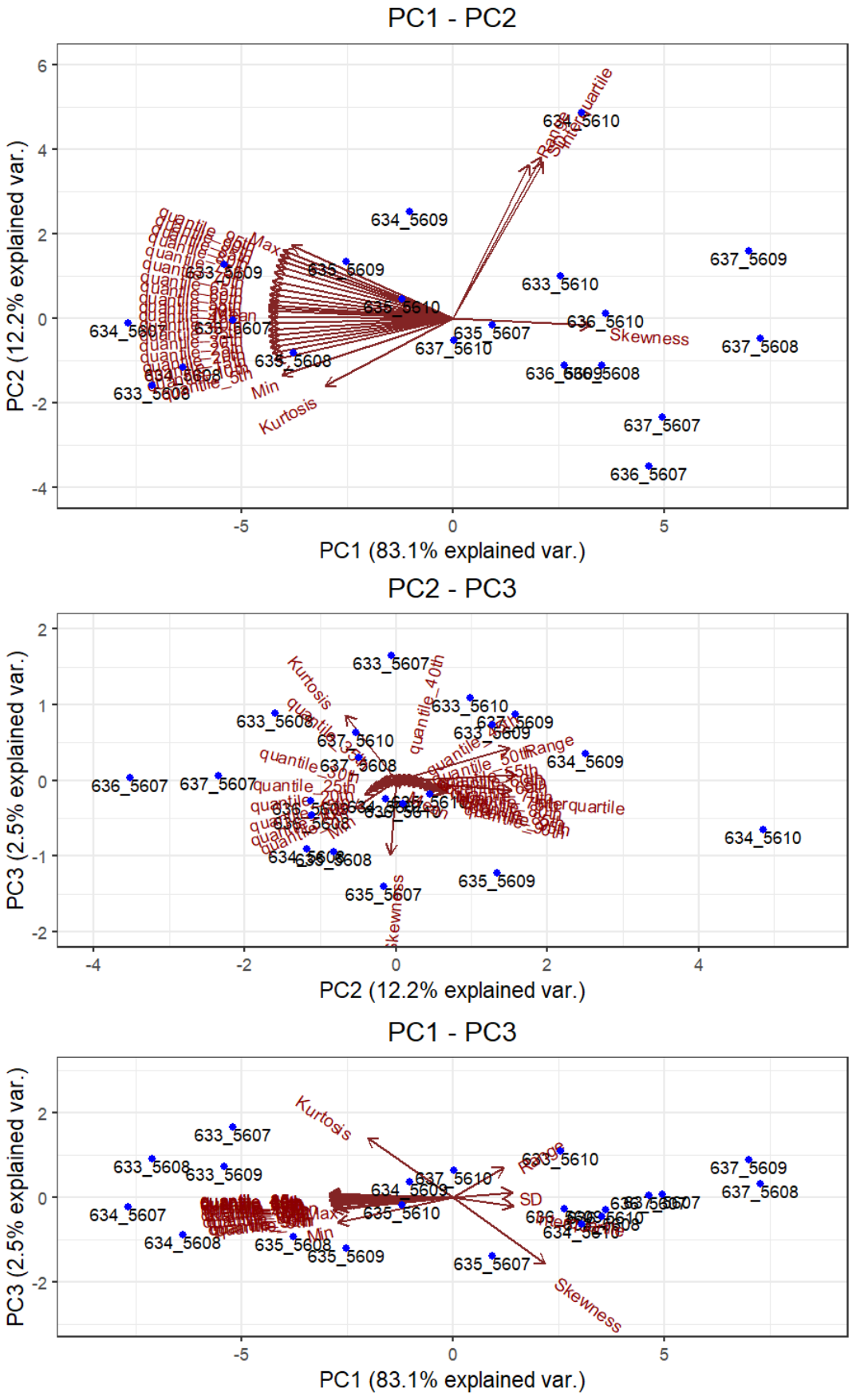

Figure 5*. Biplots of PC1-PC2 (a), PC1-PC2 (b) and PC2-PC3 (c)*

As can be seen that PC1 and PC2 already account for more than 95% of the variation, while PC3 accounted for only 2.5 % of variation which is significantly smaller than that of PC2 (Fig. 4 and 5a). Therefore, only PC1 and PC2 are enough to describe the data and were used for further clustering and classification.

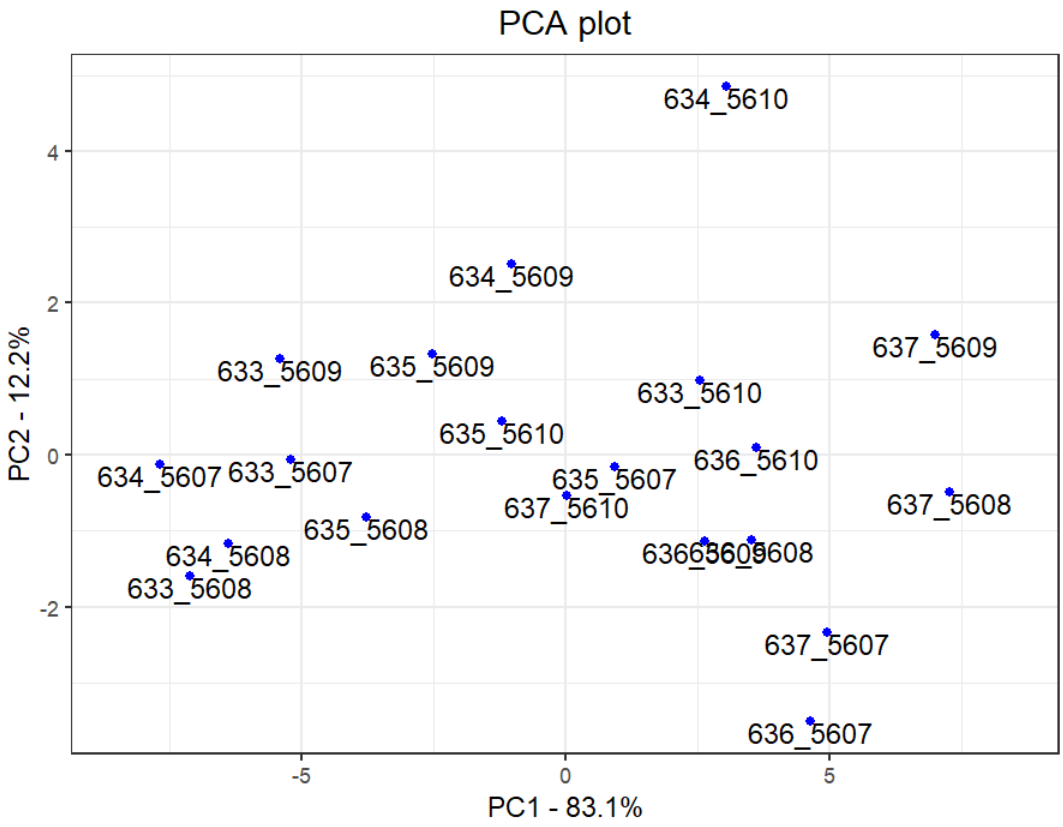

Figure 6. *PCA plot using PC1 and PC2*

Loading scores were used to see which metrics have the largest contribution to the PCs and how it affects where samples are plotted in the PCA plot (Figure 6). Metrics that push samples to the left side of the graph will have large negative values and variables that push samples to the right will have large positive values. Magnitudes of the loading scores refer to the significance of the contribution.

Table 2. *10 metrics contribute most to PC1 and PC2 and their loading score*

Rank	PC1	Loading scores	PC2	Loading scores
1	quantile_35th	-0.21069066	SD	0.10122396

2	Mean	-0.21067716	Interquartile	0.10328737
3	quantile_40th	-0.21067627	Range	0.08682688
4	quantile_30th	-0.21048452	Max	-0.18440297
5	quantile_45th	-0.21048265	quantile_95th	-0.19394014
6	quantile_50th	-0.21016096	Kurtosis	-0.14577102
7	quantile_25th	-0.20995841	quantile_90th	-0.19709412
8	quantile_55th	-0.20968995	quantile_85th	-0.19950476
9	quantile_20th	-0.20914245	Min	-0.19518627
10	quantile_60th	-0.20894821	quantile_80th	-0.20165141

3.3. K-means clustering

After dimension reduction by PCA, clustering of the selected PCs (PC1 and PC2) was performed. One of the most popular methods of clustering is the unsupervised K-means algorithm. The method requires choosing a fixed number of clusters (K) and then grouping data based on distance similarities.

Firstly, a proper number of clusters was determined by the so-called Elbow plot (Figure 7) which is similar to the Scree plot (Figure 4) in PCA. The Elbow plot shows the sum of squared distance between each point and the centroid in a cluster. The plot of the sum of squares with the K value resembles an elbow. The sum of squares values will begin to drop as the number of clusters rises. The highest sum of square value is at K = 1. When examining the graph, it is noticed that there is a point where the graph significantly changes, forming an elbow. The graph then begins to move nearly parallel to the X-axis from this point on. The best K value, or the optimal clusters, is the one that corresponds to this location. Looking at Figure 7, the number cluster of 3 may be optimal. However, for the testing reason, in this study number clusters of 2, 3, and 4 were analyzed.

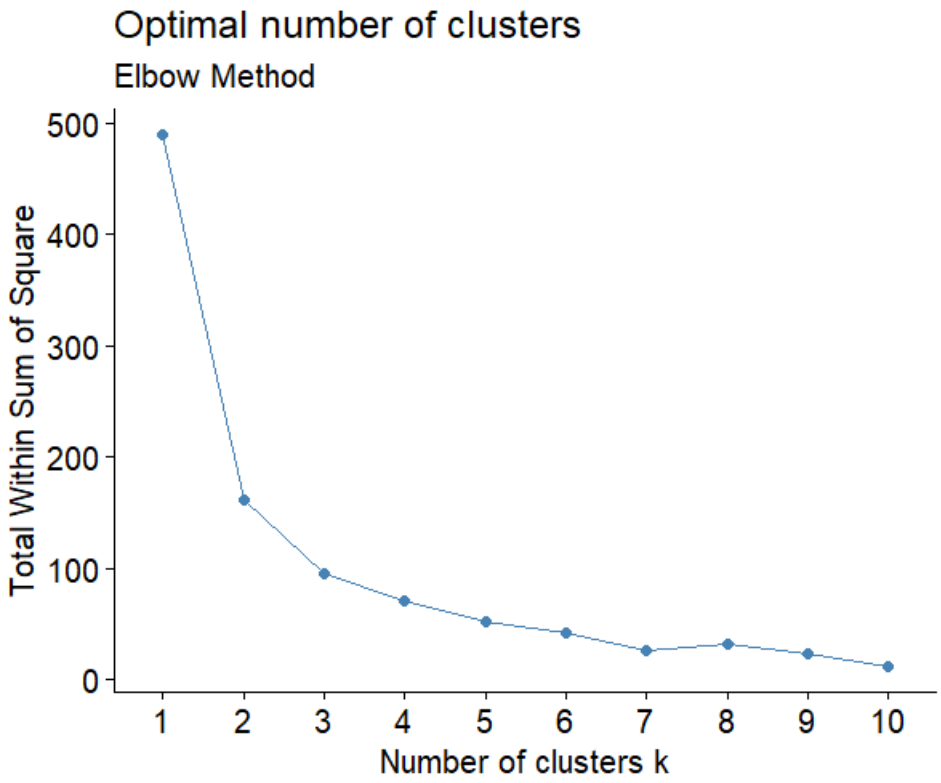

Figure 7. *Elbow plot for determining the proper number of clusters*

The algorithm randomly configures K centroid positions in the data space. After each distance calculation with all data points, the algorithm will optimize for the best position of the centroids to the middle of the clusters. The algorithm stops when no change in centroid position occurred or the best centroid position with minimum distance to all data points in the cluster was identified. Therefore, the results of K-means algorithm vary according to the randomly chosen starting centroids. In this study, K-means algorithm was applied to PC1 and PC2 data for 2, 3, and 4 clusters, with 100 sets of randomly starting centroids each (Figure 8).

It can be seen that the 3 clusters grouped by the K-means algorithm show an agreement with boxplots of the 20 tiles (Figure 9). The tiles with similarities in height distribution were grouped together.

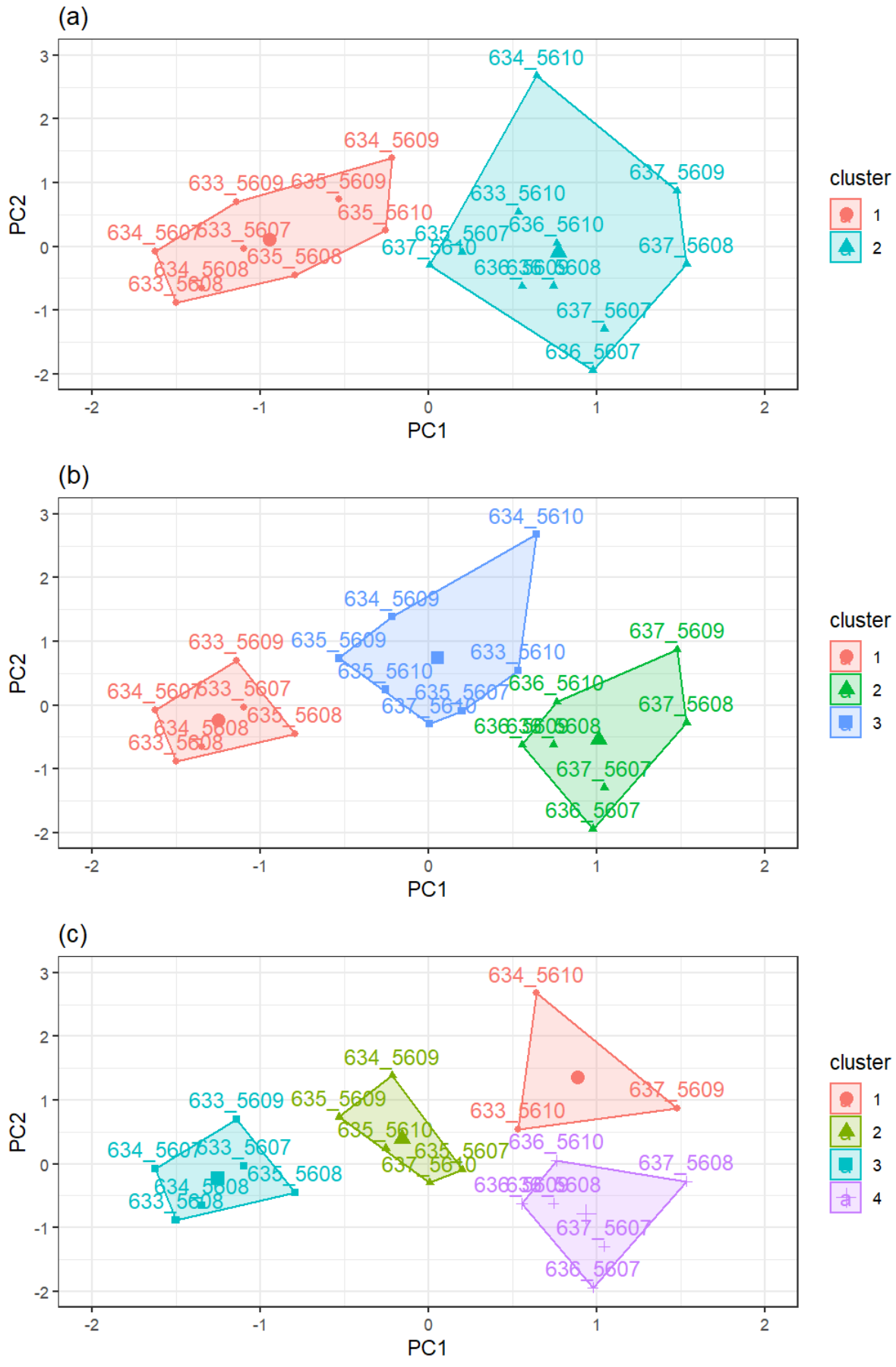

Figure 8. *K-means cluster plot with 2 clusters (a), 3 clusters (b). and 4 clusters (c)*

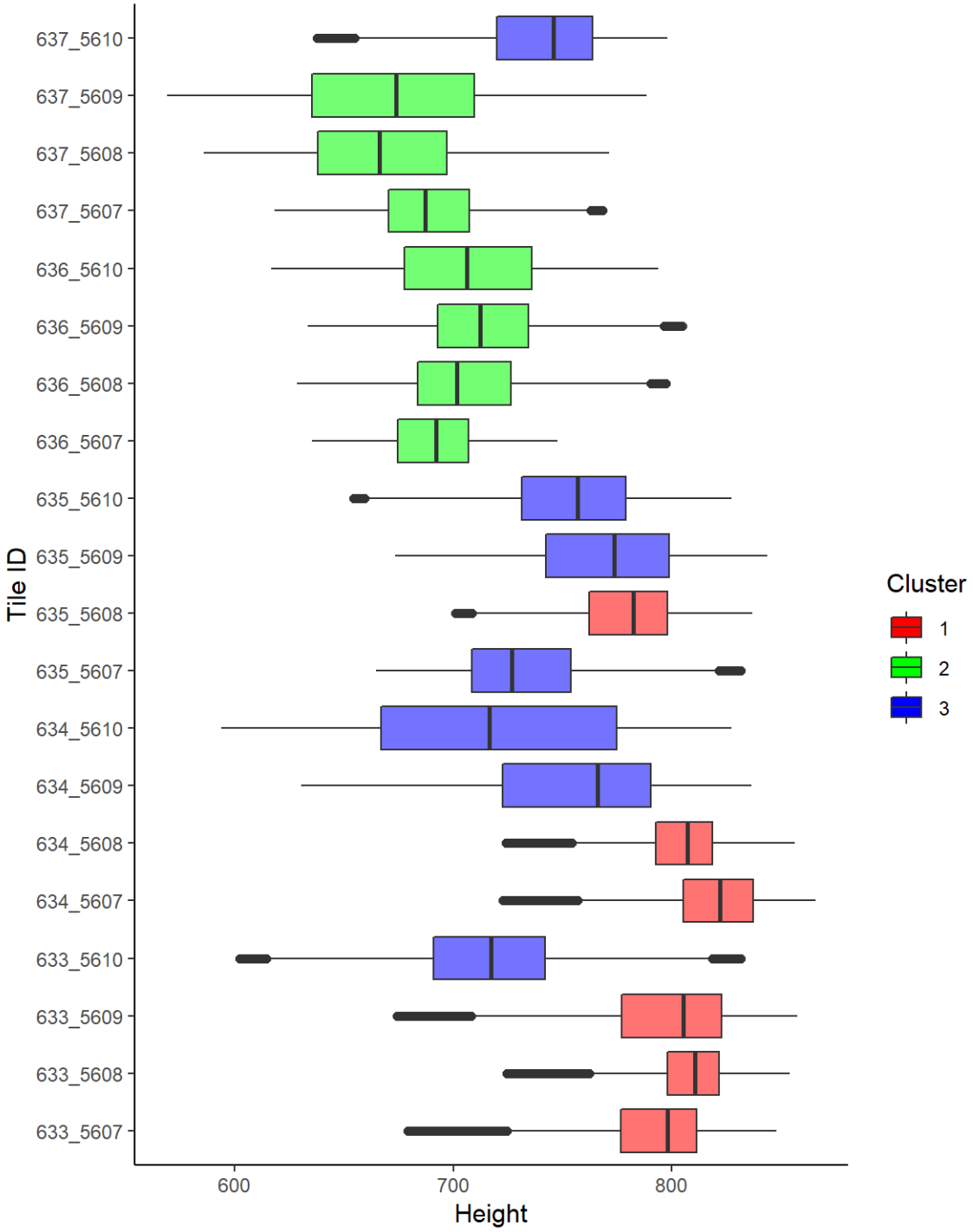

Figure 9. *Boxplots of height data from 20 laser scanning tiles grouped into 3 clusters by K-means algorithm*

3.4. Decision tree

Decision tree is a popular and powerful method for classification and regression. A decision tree is a type of tree structure that resembles a flowchart, where internal nodes present tests or yes-no questions on the attribute, branches denote the result of the question, and leaf

nodes (or terminal nodes) show class labels (GeeksforGeeks, 2017). Figure 10 presents the decision tree that classifies 18 laser scanning tiles into 3 groups based on selected PCs values. With PC1 higher than or equal to -3.1, the tiles classified as group 1 if PC1 is smaller than 2.6. Otherwise, the tiles classified as group 1 if PC2 is higher than or equal to 2.5 and conversely as group 2. Meanwhile with PC1 smaller than 3.1, the tiles classified as group 3. The 2 remaining tiles were used as test data to validate the decision tree, an accuracy of 100% was achieved from validation.

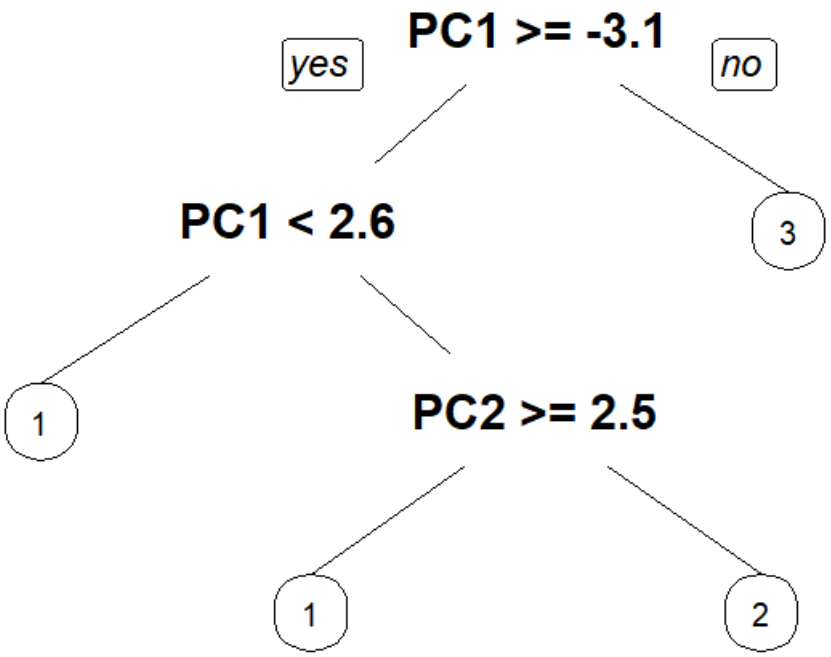

Figure 10. *Decision tree classifying 18 tiles into 3 groups from K-means clustering result*

3.5. Random forest

It can be seen that the decision tree method can suit any training data, providing 100% accuracy of any classification (section 3.4). This situation is known as “overfitting” when the decision tree performs perfectly with the training data used to create them but has low accuracy with a new dataset. It happens when a decision tree is constructed without consideration of limiting the tree size (number of leaves, splits, and depth) or pruning after training. Random forest, on the other hand, aggregation of decision trees avoids overfitting. Each decision tree in a random forest was built based on randomly selected observations and features. Since only a

subset of the sample and a few predictors are used to construct each decision tree, resulting a wide variety of trees. The variety makes a random forest more effective than an individual decision tree, especially when it comes to classifying a new dataset.

In this study, a random forest classifier was applied for the same training (18 tiles) and testing (2 tiles) dataset as the decision tree section. In random forest, tile groups were predicted based on PC1 and PC2. Figure 11 shows the error rate of random forest classification with the different numbers of trees. The red line presents the error rate when classifying tile group 1, the green line for tile group 2, the blue line for tile group 3 while the purple line for overall out-of-bag error (OOB). In general, it can be seen that the error rates decrease when the random forest has more trees, especially, the error rates stabilize after 300 trees. Therefore, 300 trees would be the optimal number of trees and would be used for further analysis.

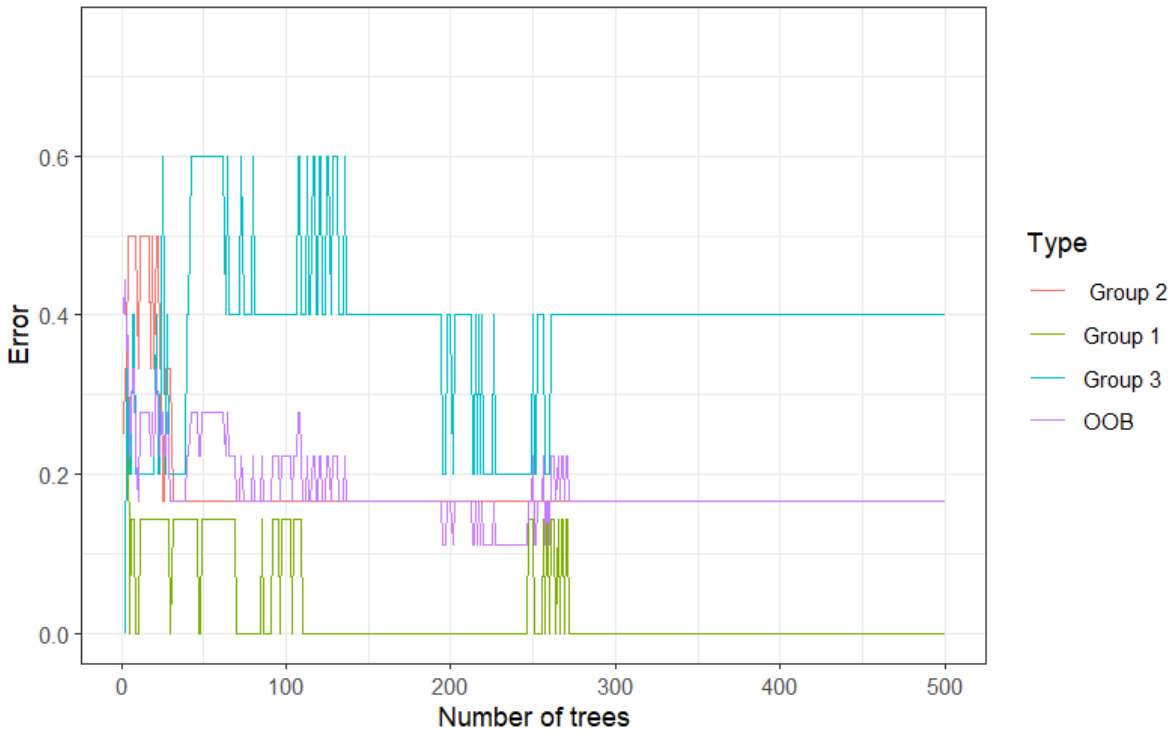

Figure 11. *Error rates of random forest classification with different numbers of tree*

The random forest of 300 trees performs better with 1 variable than 2 variables at each step when building each tree, the corresponding OOB values are 0.166 and 0.332. The Mean Decrease Accuracy plot shows the amount of accuracy the model drops if removing each

variable. The PC1 reveals more contribution to the model than PC2 (Fig. 12). The Mean Decrease Gini shows each variable contributes to the purity of the nodes and leaves. PC1 also expresses the more important impact on the purity of the random forest (Fig. 12). It is understandable as PC1 covers the most variation in the data identified by PCA, which is valuable to be used for separating data into groups. Similar to the decision tree approach, the random forest also acquired 100% accuracy with the 2 testing tiles.

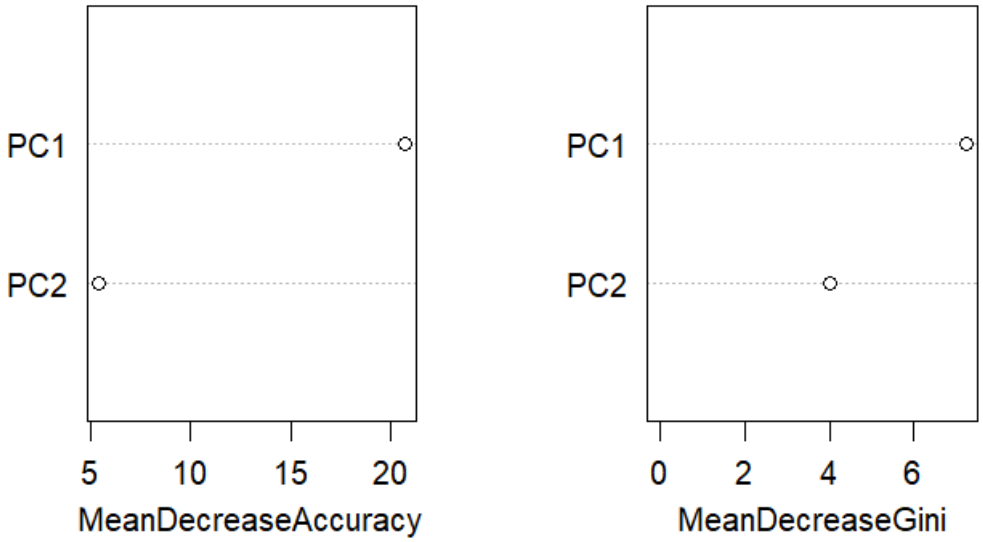

Figure 12. *Variable importance plot for the random forest*

APPENDIX I. DISCRIPTION OF THE LIDAR DATA

Table 3. *Information of the LiDAR data*

Tile ID	Collection date	Area
633_5607	2019-04	227_2019
633_5608	2019-04	227_2019
633_5609	2019-04	227_2019
633_5610	2015-03	207_2015
634_5607	2019-04	227_2019
634_5608	2019-04	227_2019
634_5609	2019-04	227_2019
634_5610	2015-03	207_2015
635_5607	2019-04	227_2019
635_5608	2019-04	227_2019
635_5609	2019-04	227_2019
635_5610	2015-03	207_2015
636_5607	2019-04	227_2019
636_5608	2019-04	227_2019
636_5609	2019-04	227_2019
636_5610	2015-03	207_2015
637_5607	2019-04	227_2019
637_5608	2019-04	227_2019
637_5609	2019-04	227_2019
637_5610	2015-03	207_2015

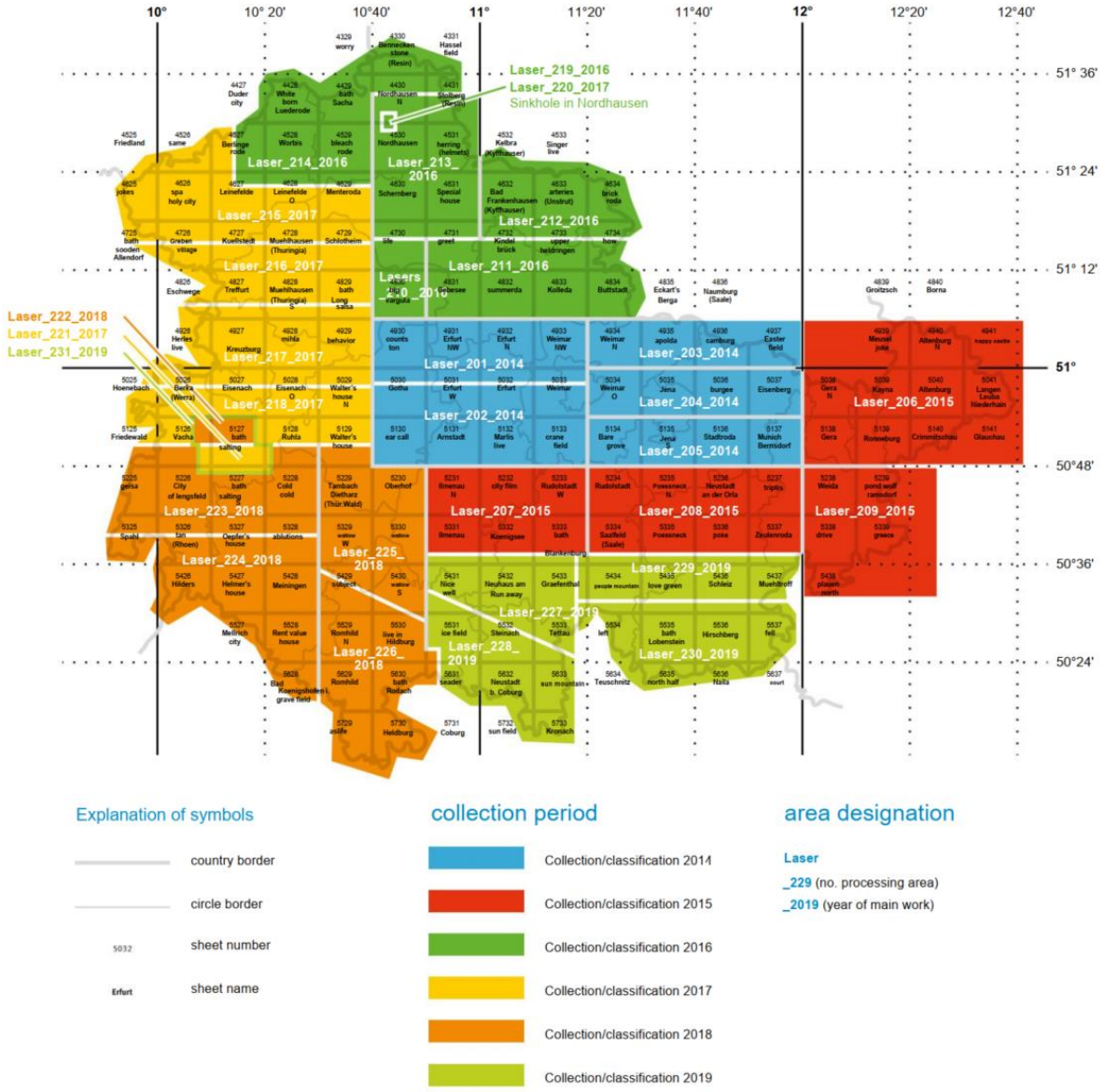

Figure 13. *Cyclic acquisition of airborne laser scanning in Thuringia.*

REFERENCES

GeeksforGeeks. (2017). *Decision Tree - GeeksforGeeks*.

https://www.geeksforgeeks.org/decision-tree/

Wikipedia (Ed.). (2022). *Thuringia*.

https://en.wikipedia.org/w/index.php?title=Thuringia&oldid=1098601799

Worlddata.info. (2022, July 19). *Climate: Thuringia, Germany*.

https://www.worlddata.info/europe/germany/climate-thuringia.php

Zhang, Z. (2017). *Environmental data analysis: Methods and applications*. De Gruyter.

https://doi.org/10.1515/9783110424904