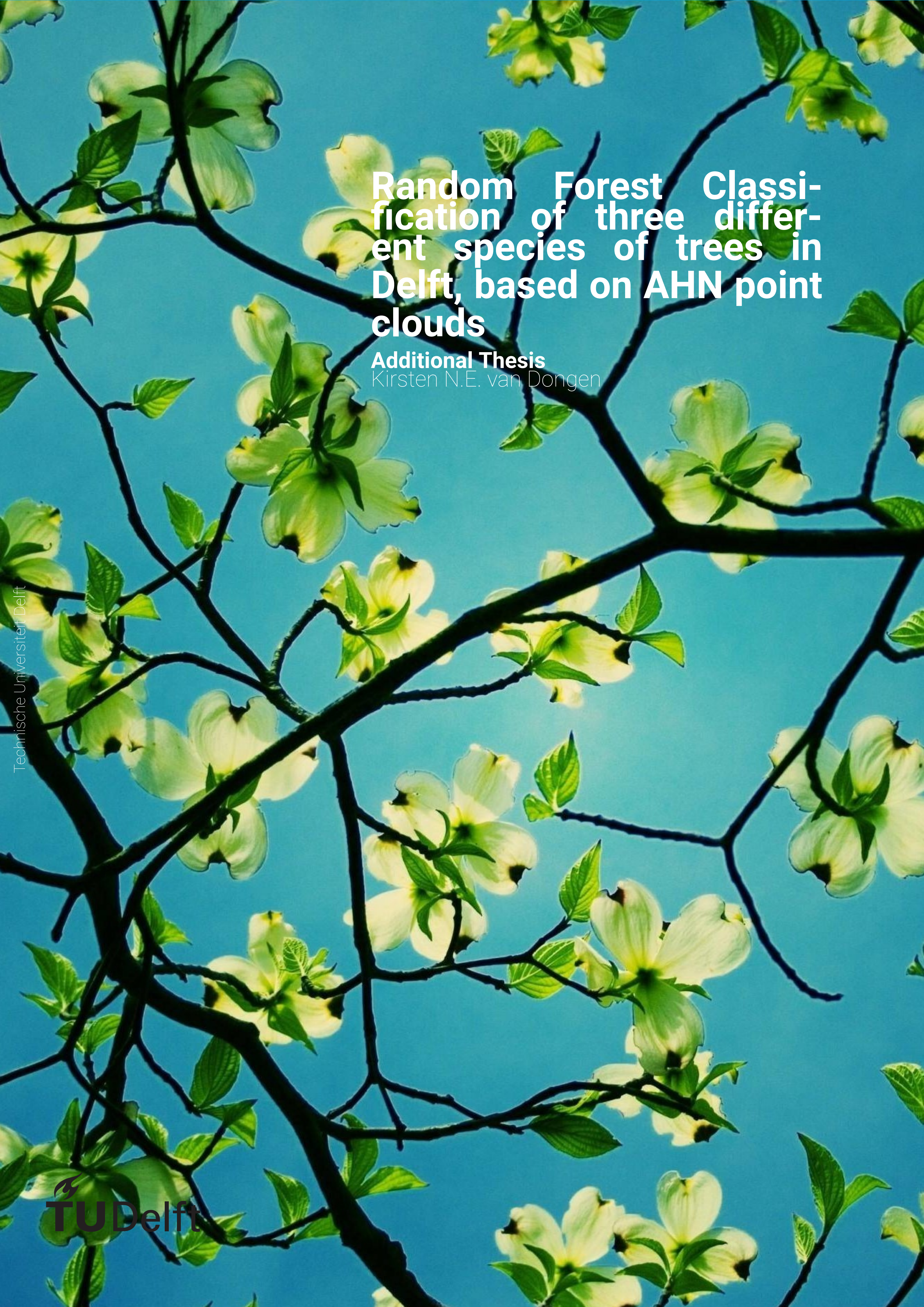


Random Forest Classification of three different species of trees in Delft, based on AHN point clouds

Additional Thesis

Kirsten N.E. van Dongen

Technische Universiteit Delft

Random Forest Classification of three different species of trees in Delft, based on AHN point clouds

Additional Thesis

by

Kirsten N.E. van Dongen

Student number:	4036662	
Project duration:	September, 2018 – October, 2018	
Thesis committee:	Dr. R. C. Lindenbergh,	TU Delft, supervisor
	Dr. Liangliang Nan,	TU Delft (Faculty of Architecture)

Abstract

Trees are an important aspect of the world around us, and play a sufficient role in our daily lives. They contribute to human health and well-being in various ways. Tree inventory and monitoring are of great interest for biomass estimations and changes in the purifying effect on the air. It is a very time consuming and cost inefficient way to check every tree in and around a city or town, therefore there is further research required in the use of AHN data. Together with the “tree information data set” form the municipality of Delft, the location and the corresponding point cloud of tree different species of trees are selected. For the species of interest, Aesculus Hippocastanum, Acer Saccharinum and Platanus x Hispanica, different characteristics are determined. In this research six different characteristics are estimated; Height, Trunk Height, Normalized Trunk Height, Canopy Projected Area, Normalized Canopy Projected Area, Ratio of Diameters, Normalized Ratio of Diameter, Centre of Gravity and at least the Normalized Centre of Gravity. These characteristics are used as features for the Random Forest Classification, Consequently the Confusion Matrix is used as performance measurement. The results of a test of 30 pointclouds, per species of interest, show that the Random Forest Classification is able to classify individual trees. However, these three different species cannot by sufficiently classified using clustering.

Preface

Before you lies the additional thesis *Random Forest Classification of three different species of trees in Delft, based on AHN point clouds.* It has been written to fulfill the additional thesis requirements of the master study *Geoscience and Remote Sensing* at the Delft University of Technology. The main focus of this research is on data processing of laser scanned data, and the classification of three different species of trees. I was engaged in researching and writing this thesis in September and October 2018. The final presentation, for the supervisors and interested public, was on the 29th of October 2018.

During my master, I followed the course *3D Surveying of Civil and Offshore Infrastructure (2017/2018).* The course contained of a project, whereby I, and two follow students, looked at the AHN data in Delft and tried to construct the shadow of a tree during several moments of the day [17]. This project really appealed to me, it was intuitive and the results where very straight forward. Therefore I wanted to do some further research in point cloud processing and discover further applications of the tree data set we created. In consultation with Roderik Lindenbergh, we came across this subject for an addition thesis.

I gained a lot of satisfaction during the project. I became interested in the different aspects of data processing and even in the mathematics behind it. Especially the different methods of classification appealed to me. A downside of this project, for me, where the multiple assumptions that where necessary to made. I seem to have a preference for more exact research and more exact approaches. Although it was sometimes frustrating for this project, it taught me that I should avoid this -as far as possible- for my final master thesis and that I am able to overcome this kind of frustrations.

I would like to thank Roderik for supervising and supporting me through the whole process. He asked many critical questions and was able to answer many of my (hopefully critical) questions. Furthermore I would like to thank the "green department" of the municipality Delft, for telling me more about the database they provided. I benefited from debating issues with my friends and family. Corné Verhoeven deserves a particular note of thanks; your motivation and the time you take to help me with my programming part served me well.

I hope you enjoy reading this report.

Kirsten N.E. van Dongen
Delft, October 2018

Contents

1

Introduction

1.1. Study Area

Delft is located in the West of the Netherlands. It is enclosed between the well-known cities of The Hague and Rotterdam. The city or Delft is known for it's historic centre with idyllic canals and houses. Delft is somewhat famous for the Dutch painter Johannes Vermeer. Delft is home to Delft University of Technology (TU Delft). In the East of Delft lies the "Delftse Hout", it is a nature- and recreation- area. There is a lake in the middle of this small forest, surrounded by narrow beaches, a restaurant and community gardens. Inside the city, there are several smaller town parks, like "Nieuwe Plantage","Wilhelminapark","Mekelpark" and other. Furthermore there is the Botanical Garden of the University and a glasshouse in the Delftse Hout.

1.2. Relevance of the Research

Trees are an important aspect of the world around us, and play a sufficient role in our daily lives. They contribute to human health and well-being in various ways. All trees -and other vegetation- convert carbon dioxide into oxygen, which ensures healthy and breathable air. In addition to this, the presence of trees also provides space where relaxation and sport activities can be achieved. Not all threes have the same purifying effect on the air, and therefore it is important to know which three is located where [9]. It is a very time consuming and cost inefficient way to check every tree in and around a city or town. And not only the species of the trees is necessary information, but also the height, canopy projection, trunk length are important characteristics for tree management [19] [9]. Manually measuring all these features is inefficient regarding labour hours and the cost, therefore a new method should be developed.

1.3. Acknowledgement of previous work

The identification of trees is -so far- done using "determination steps" developed by biologists. The determination table asks many questions, and answers will lead to the species of the trees. Question like the shape of the leaves, the colour of the leaves and the trunk etc. are common in this determination process. However, this process is very labour intensive and therefore quite expensive. Other methods developed to determine the species of trees is done by Mobile Laser Mapping [19]. It is still labour intensive, because the mobile laser scanner needs to be placed next to the trees of interest. In order to make the process even more efficient, another method of classification should be developed.

1.4. Problem statement and project description

This research is an extension of the study done by Jinhu Wang, in collaboration with Roderik Lindenbergh [19]. In this additional thesis one does not use Mobile Laser Scanned trees, but AHN data. The advantage of this method is that the AHN data set is free to download from the Dutch government website. Therefore, it becomes even less labour intensive to make a classification. To make this study a good comparison with Jinhu Wang's research, the same species of interest were chosen.

1.5. Research Question

The above mentioned objectives are transferred to a research question. In this report this question will be answered and described. The question addressed in this thesis is: **Is it possible to identify the *Aesculus Hippocastanum*, *Acer Saccharinum* and *Platanus x Hispanica*, based on AHN3 data, in Delft?** The relevant sub-questions are;

- **Is it possible to determine different features from AHN point cloud data?**
- **Which features of trees can be used to classify these trees of interest?**
- **Can classification be done using the supervised Random Forest Classification?**
- **Is it possible to classify the tree different species of trees with clustering?**
- **How do some of the wrongly classified trees look like? And why are they wrongly classified?**

1.6. Report Outline

This report consist of six chapters. Chapter 2 gives an introduction about the different data sets used for this research. First a short introduction about trees in general, followed by information about the data set -provided by the municipality of Delft- is given. Afterwards the AHN3 data set is introduced and explained. In Chapter 3 the methodology is presented. A step-by-step approach of answering the research question is described. In the following chapter, Chapter 4, the results are presented in detail. The conclusion is described in Chapter 6, in Chapter 7 the discussions are described and recommendations for further research are done.

2
Background Information

In this chapter the different kinds of required data is described and necessary background information is provided. First the different characteristics of trees are discussed, and the data sets with the locations of different trees is introduced. Information about the three species of interest are provided and a short explanation about the point cloud data is given afterwards. The hereby associated laser altimeter is explained in the last section.

2.1. Introduction to Trees

There is no universally agreed definition of a *tree*, even in the science of botany. Most of the time a tree is defined as a plant, that lives longer than 2 years, with an elongated stem or trunk (mostly wood) and supporting branches. Most of the species have leaves or needles. [9]

2.1.1. Characteristics of a Tree

A tree has different features that distinguish a tree from a plant, and characteristics that distinguish different kinds of trees. In this paragraph the different features are described, focusing on the trees that can grow in the climate of the Netherlands.

Trunk

The trunk of a tree is defined as the main woody stem of a tree. From this structure, smaller branches arise. Inside the trunk are the veins that connect the roots with the canopy. Inside the veins transport of nutrients and water from the roots to the canopy takes place. Vise versa, the waste is transported from the canopy to the roots [9]. This definition of a trunk is inadequate, because it is too vague where the main woody stem is a trunk and where is becomes bigger branches, and thus where the canopy starts. In figure 2.1 one can see multiple opportunities to define the trunk of two different trees.

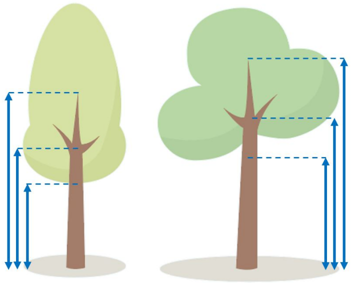


Figure 2.1: Different concepts for the trunk

Canopy

The canopy is defined as the uppermost branches of the trees in a forest, which forms a more or less continuous layer [9]. The canopy depends on the species of the tree and the season. The Netherlands is subject to four different seasons during one year. In the spring, the leaves of deciduous trees start developing and the canopy grows. In the autumn the leaves fall, due to lack of nutrients. The canopy is at maximum size during the summer, and minimum during the winter. Conifers has a constant canopy size over the year [11].

The shape of the canopy can differ very much. There are several aspects that influence this shape. One of the most important ones is the presence of obstacles in the direction of growth. One can think of a building next to the tree, resulting in a very asymmetrical growth. A lack of enough incidental sunlight can interfere with a good development of the canopy [10]. For example in a thick forest, the smaller trees stay small, because of lack of space and sunlight [9] [11]. The definition of the canopy is inadequate, it is too vague what the uppermost branches are and the alternations in amount of leaves during different seasons. In figure 2.2 the different conceptions for the canopy are shown.

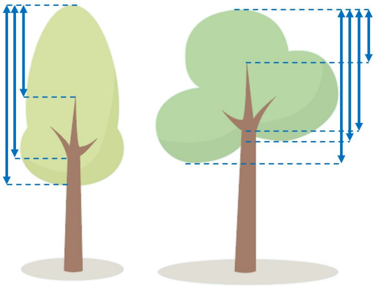


Figure 2.2: Different concepts for the canopy

Roots

The definition of the roots is that it is the part of a plant which attaches it to the ground and nourishes the rest of the plant via numerous underground branches and fibres. However not the entire root is under the ground, especially by older and bigger trees [9] [16].

Centre of Gravity

The centre of gravity is an imaginary point in the tree, which may be considered as the point from which the gravity engages. This is not only the case for trees, but for all objects [8]. The centre of gravity gives information about the direction of growth [18]. This point depends on the canopy, thus it depends on the seasonal cycles as well. In figure 2.3, three different centres of gravity examples are shown. The arrow represents the "vector of growth", between the beginning of the trunk at the ground and the centre of gravity.

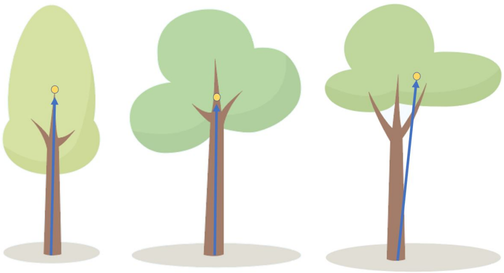
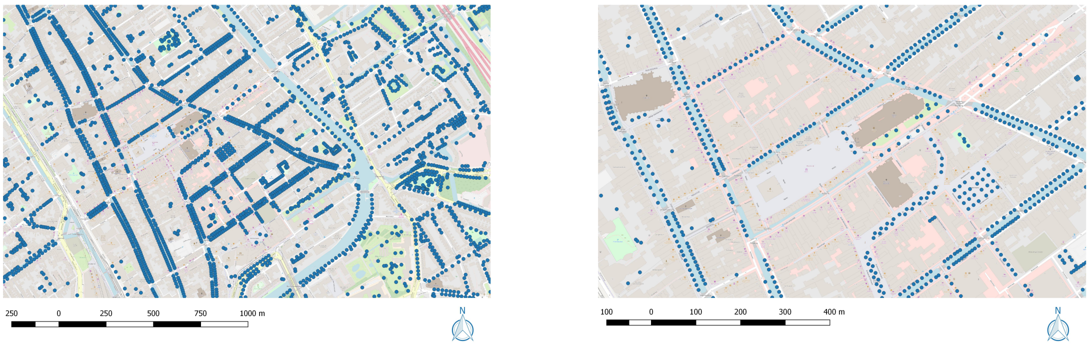


Figure 2.3: Centre of Gravity for different shapes of trees

2.1.2. Trees in the Municipality Delft

Delft municipality has about 38,000 trees under management. It mainly concerns trees in public space; along roads, in parks, and in public gardens. The data on monumental trees in private ownership are also included. The information on these trees is available as open data [4]. The data contains information about the x- and y-coordinates (in Rijksdriehoek coordinate system), health, species (Dutch and Latin names), trunk diameter approximation, estimated height of the tree and whether the tree is detached in or in the direct surrounding of another tree or building. In figure 2.4 one can see the trees that are located in Delft. In the left figure (figure 2.4a) one can see the trees in the city centre of Delft and the right image, (figure 2.4b) shows the market square with the locations of trees.



(a) Trees in the city centre of Delft

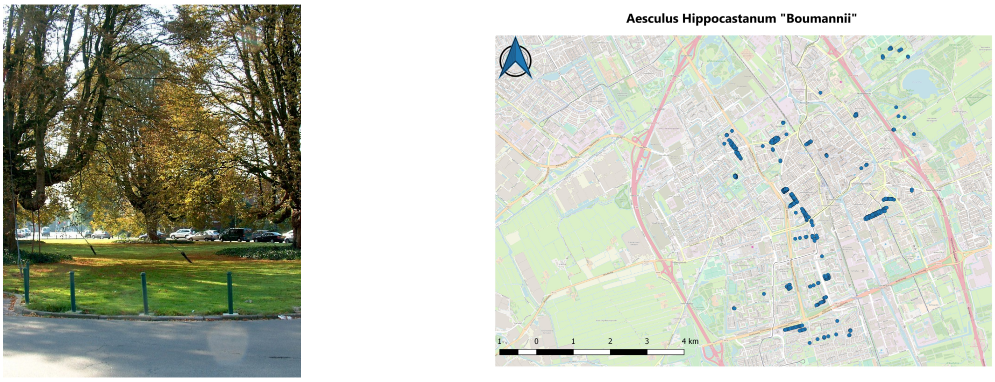
(b) Trees at and around the market square in Delft

Figure 2.4: Location of thees under management of municipality Delft

2.1.3. Species of Interest

This research focuses on three different species of trees, which are often recurring in Delft. They have some characteristics that are very different and some features are corresponding to each other. In this section, the three different species are described.

Aesculus Hippocastanum "Baumannii"
The *Aesculus Hippocastanum "Baumannii"*, is called *Dubbelbloemige Paardenkastanje* in Dutch. It is a white flowering chestnut that grows about 25 meters high. The *Aesculus Hippocastanum* gets a spherical crown, that is the highest part of the tree where the leaves and the branches are. The leaf is dark green and hand-shaped, it can get 20 cm tall. The flowers of the *Aesculus Hippocastanum* are white/pink and appear in May/June, the flowers do not bear fruit [7]. In figure 2.5a one can see a photo of the chestnut in Delft, taken in September. This *Aesculus Hippocastanum* makes little demands on the soil, and therefor grows in almost all places. It is especially a suitable tree for wide avenues and streets, but also as a solitary tree in a large garden or in a park [7]. In Delft they occur 436 times, in figure 2.5b one can see the location of these trees: [4]



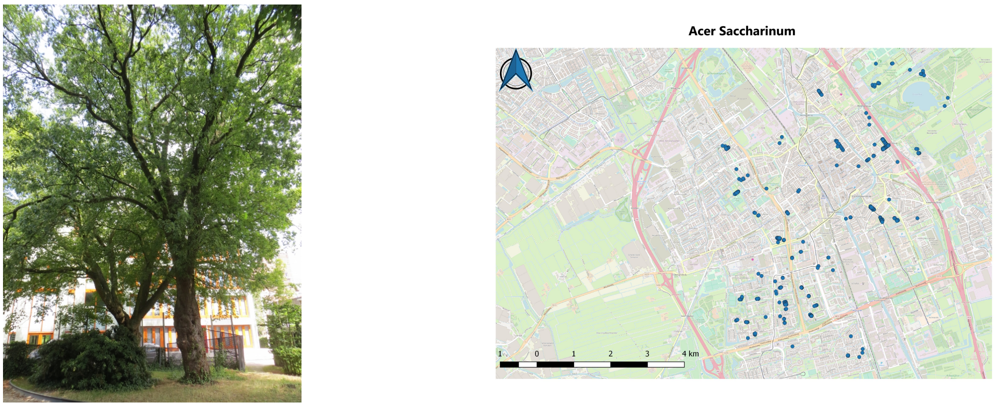
(a) *Aesculus Hippocastanum* in Delft

(b) Location of the *Aesculus Hippocastanum* in Delft

Figure 2.5: *Aesculus Hippocastanum* "Baumannii"

Acer Saccharinum
The *Acer Saccharinum* is called the *Zilver Esdoorn* in Dutch. It is a fast-growing maple tree (to 25 meter) that develops a broad fan-shaped crown. The tree has five-lobed, deeply incised leaves whose underside is white

and the upside green. The twigs are thin and flexible, with striking red, crosswise opposite buds. The flowers are red/green and quite small and appear in February/March [1]. In figure 2.6a one can see a photo of a *Acer Saccharinum* in Delft.
It is a tree that places few demands on the growing location. The tree thrives on both wet and dry soils and hardening is well tolerated. For this reason, the silver maple is often used as a street or avenue tree. The broad crown requires a lot of space [1]. In figure 2.6b the locations of the *Acer Saccharinum* are shown in Delft.They occur 304 times in Delft and surroundings [4].

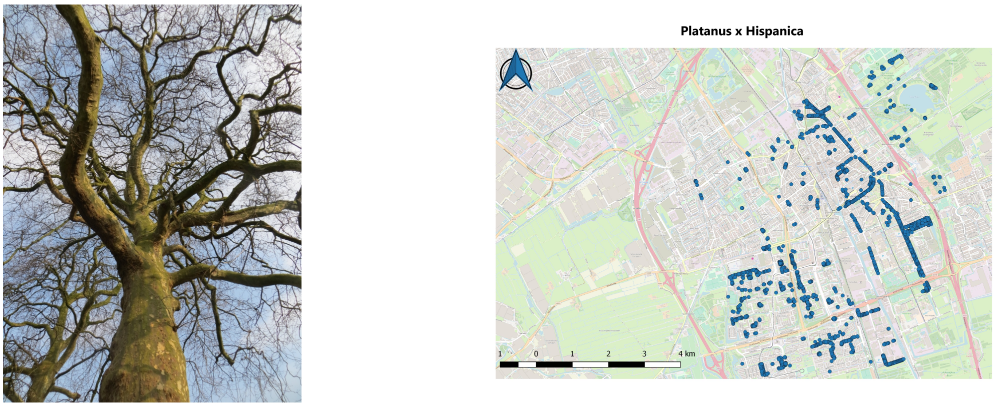


(a) *Acer Saccharinum* in Delft

(b) Location of the *Acer Saccharinum* in Delft

Figure 2.6: *Acer Saccharinum*

Platanus x Hispanica
The *Platanus x Hispanica* is called the *Leiplataan* in Dutch. It is a tree with a long, straight trunk. Characteristic is the peeling bark, with a light yellow/green surface and green/gray darker parts. The broad, hand-shaped 12-25 cm leaves, have 3,4 or 5 pointed serrated lobes and is slightly hairy on the underside. The tree can get 35 meters high and has a spherical, half open crown. The flowers are green/yellow spheres. In figure 2.7a one can see a photo of a *Platanus x Hispanica* in Delft.
This three has few demands on the growing location. The fallen leaf digests badly and therefore remains around the tree for a long period of time [13]. In figure 2.7b one can see the locations of the *Platanus x Hispanica* in Delft. In Delft, they occur 2180 times. [4]



(a) Platanus x Hispanica in Delft

(b) Location of the Platanus x Hispanica in Delft

Figure 2.7: Platanus x Hispanica

2.2. AHN Data

For this research an existing point cloud containing the elevation in the city of Delft was used. The *Actueel Hoogtebestand Nederland (Current Altitude Netherlands)*, further referred to as AHN, is a file with detailed elevation data for a big part of the Netherlands. For almost every squared meter in the Netherlands, the elevation in relation to the *Normaal Amsterdams Peil (Normal Amsterdam Level)*, further called NAP, is known. The data collected for the AHN is done by laser altimetry. More information about laser altimetry in the next section. A number of products of the AHN data have been made, which can roughly be dived into two categories: grids and 3D point clouds. Both are projected in the WGS84 coordinate system. The AHN Digital Terrain Model (raster) is intended as ground level file, where all points classified as "ground level" are re-sampled to a grid, based on a Squared Inverse Distance Weighted Interpolation. Points classified in other classes (buildings, bridges, water, trees, roads etc) are not used in the re-sampling.

In the AHN Digital Surface Model, all points except those classified as "water" are re-sampled into a grid based on a Squared Inverse Distance Weighting method. This point cloud data is a file where a classification process is applied to the individual points. [4] [6]

In this research the point cloud data from the city of Delft was used. This point cloud is a .LAZ file where a classification is applied to all the individual points. Each point is assigned to one of the following classes: ground level, buildings, water, vegetation, art work or other. In addition, extra attributes are included per point. [6]
Delft is divided in two AHN data storage files, these are divided in smaller tiles to make the computations faster and more manageable. An example of the point cloud data is shown in figure 2.8. In blue the lower elevation area's are shown, green the somewhat higher area's and the top of the church is orange/red.

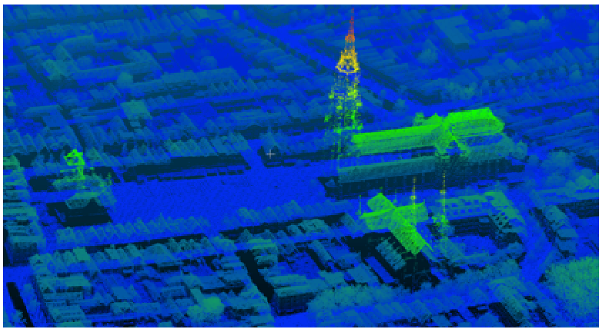


Figure 2.8: Point Cloud data of the Market Square with the New Church in the city centre of Delft

2.2.1. Laser Altimetry

Laser Altimetry is an often used method to measure the elevation of an area from an aircraft or helicopter. The distance from the plane to the Earth's surface is determined by measuring the travel time of a flash of infrared laser light. The transmitter emits a laser pulse, which travels to the surface, where they are reflected or absorbed. Part of the reflected radiation returns to the receiver, and the travel time of the pulse is measured. The distance between the plane and the Earth's surface is then calculated using the formula:

$$d = c * t * \frac{1}{2} \tag{2.1}$$

Where the d stands for the distance (in meter) between the plane and the surface, c is the speed of light (300 * 10^6 ms^{-1}) and t the total travel time (s). [14] [15]. In figure 2.9 a simplified image of the pulse emission and reflection is shown. On the left side of the figure, the situation for a flat ground surface is shown. On the right side the situation with an uneven ground surface. The travel time of the pulse is shorter, because the distance between the plane and the tree is smaller.

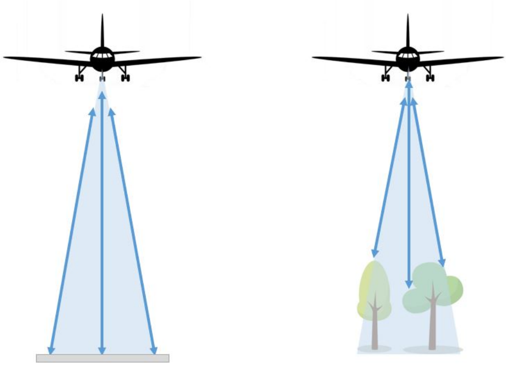


Figure 2.9: Simplified representation of the working of laser altimetry

As becomes clear the travel path for each pulse differs from the other pulses. The angle of reflection and the angle of income at the receiver vary per pulse. This makes the determination of the exact height more difficult [14] [15]. In order to determine the exact coordinates of any surface spot that was hit by a laser pulse, it is necessary to know two more parameters: the location of the aircraft from which the measurement was made, and the direction in which the laser altimeter was "looking'. These values are usually obtained through GPS-receivers and an Inertial Measurement Unit (IMU). This unit measures a body's specific force, angular rate and sometimes the magnetic field. In the case of data collection for AHN, it measures the angle of the plane which influence the laser's propagation direction. [14] [15]. However, how exactly this is done and the impact on the accuracy, is out of the scope of this research.

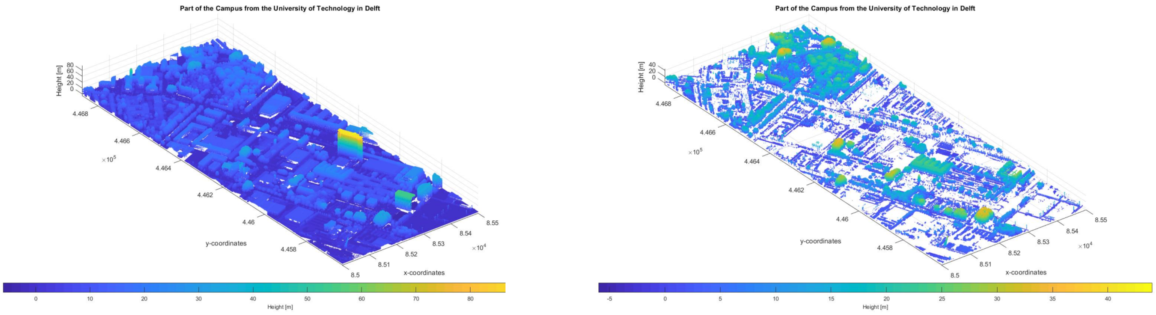
A digital image is created with the collected and calculated data. The surface spots in the original elevation data are stored as a point cloud data set. For the AHN, there are an average of 8 measurements per squared meter. The point cloud data is usually not distributed evenly over the whole area of interest. The effects of this phenomenon on this research are presented in the chapter *Discussion and Recommendations*.

3 Methodology

In this chapter the methodology is described. It starts with the preparation of the different data sets, and it ends at the classification of the three different trees of interest. All the different steps and assumptions made during the process of classification, are outlined.

3.1. Data input

As mentioned in the previous chapter, there are two different data sets that are used. The first one is the AHN data, this is a point cloud that covers a big part of the Netherlands. For this research the study area is Delft, therefore the two tiles that cover Delft and the area surrounding this city is taken into use. It was convenient to download the point cloud data as al $.laz$ file [4]. Using the laz2las tool [5] the $.laz$ files are converted into $.las$ files, because the programming environment *Matlab* can handle this kind of files. However the two point cloud files of Delft consume a lot of computer memory and makes calculations a time consuming process. Therefore one of the tiles is cut into 100 smaller tiles, whereby 2 tiles are used to do the first steps of programming. One can read the files in the programming environment *Matlab* using the command *lasdata('filename')*. As told in the previous chapter, every point in the AHN point cloud data set has a label. This research is focused on vegetation, for efficiency reasons, only the vegetation points are taken into account and all the other points are removed. In figure 3.1 two small tiles of Delft in AHN data are presented. On the left side, figure 3.1a, the AHN data with all the classes are shown and on the right side, figure 3.1b only the points with label "vegetation".


Part of the Campus from the University of Technology in Delft
Height [m]
y-coordinates
x-coordinates
×10^5
×10^4
Height [m]
Part of the Campus from the University of Technology in Delft
Height [m]
y-coordinates
x-coordinates
×10^5
×10^4
Height [m]

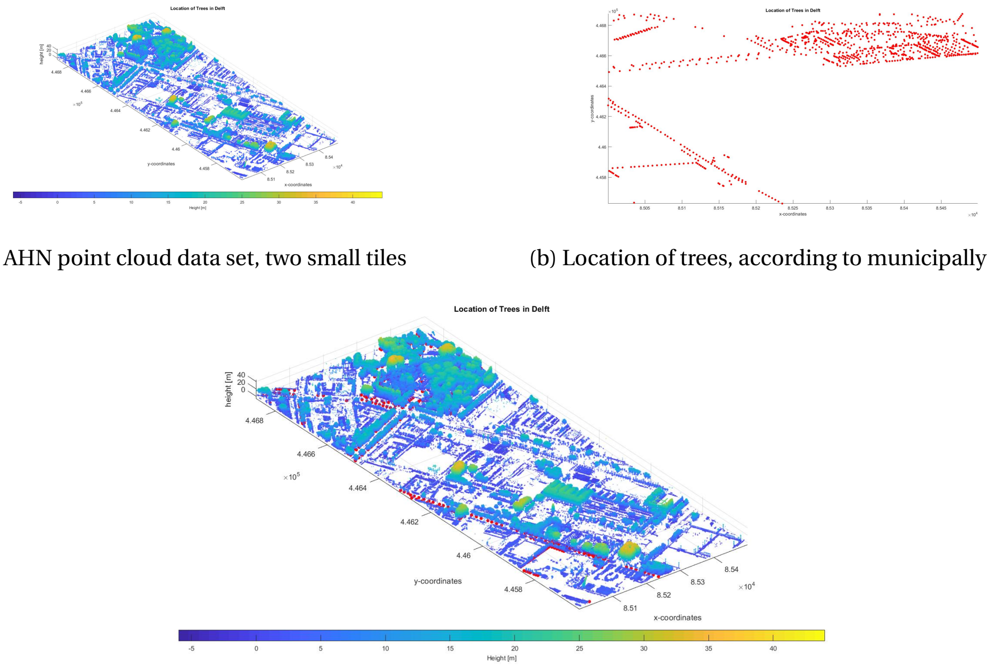
(a) AHN3 data with all the classifications

(b) AHN3 data with only the vegetation classification

Figure 3.1: AHN3 example, at the University of Technology in Delft

The second data set is the "tree information data set" from the municipality Delft, as described in the previous chapter. The information needed for this research is the location in x- and y- coordinates. It is convenient to download the data as a shapefile, instead of an excel file [4]. This way the coordinates are in more detail, such that overlap between points does not occur. The downloaded shapefile is in WGS84, but to combine it with the point cloud data it needs to be in RD-coordinates (Rijksdriehoek coordinatenstelsel). This is done in QGIS, or can be done in every other GIS program. The shapefile is opened in QGIS and converted to the RD coordinate system. The attribute table, with all the necessary information, is saved as *ESRI Shapefile*, this way the details of the coordinates are saved. One can read the data in the programming environment *Matlab* using the command *shaperead('filename')*

The AHN data and the location of the tree's are combined. Both of the data sets are in the RD coordinate system, and can be combined. In figure 3.2 the combination of the two different data sets is shown.


Location of Trees in Delft
Location of Trees in Delft
Location of Trees in Delft
height [m]
y-coordinates
x-coordinates
Height [m]
×10^5
×10^4

(a) AHN point cloud data set, two small tiles

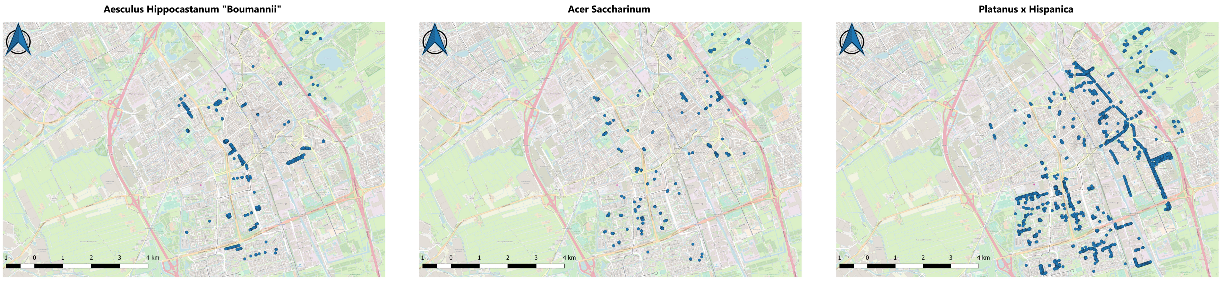
(b) Location of trees, according to municipally Delft

(c) Location of the Trees, combined with the AHN point cloud data

Figure 3.2: Location of thees under management of municipality Delft

3.2. Selection of the individual trees

After refining and combining the two data sets, the next step is to automatically extract different trees. Therefore the "tree information data set" is used. The species of all the trees are known and the corresponding locations. The location of the trees of interest are shown in figure 3.3.



(a) Location of *Aesullus Hippocastanum*

(b) Location of *Acer Saccharinum*

(c) Location of *Platanus x Hispanica*

Figure 3.3: Location of thees of interest, in Delft

For every tree of interest, whereby the precise locations are known, the corresponding AHN data needs to be determined. The approach was to create a buffer around the x- and y-coordinates of the trees [17]. This buffer depends on a certain radius, which therefore had to be estimated.

To determine the radius, preliminary information about the trees in Delft is used and a small fieldwork is

provided. During this fieldwork the trees are studied and an approximation of the canopy width is made. The final radius is based on information provided during this fieldwork and trial and error in *Matlab*. This radius is determined for each species: *Aesculus Hippocastanum* has a radius of 5,5 meter. *Acer Saccharinum* and the *Platanus x Hispanica* both have a radius of 5 meter. [7] [1] [13]
In figure 3.4 one can see the individual buffers per defined (x,y) location, representing the trees. The buffer is in 3D, but only the 2D visualization is shown in the figure to be able to see the selected points for the individual trees.

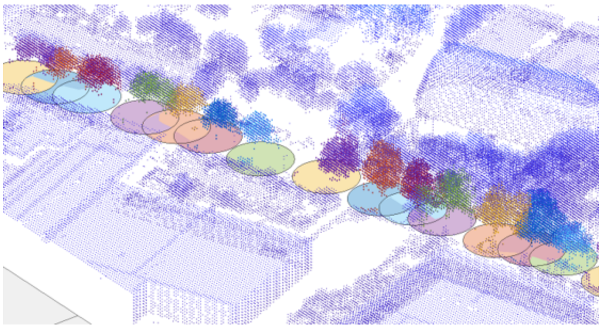
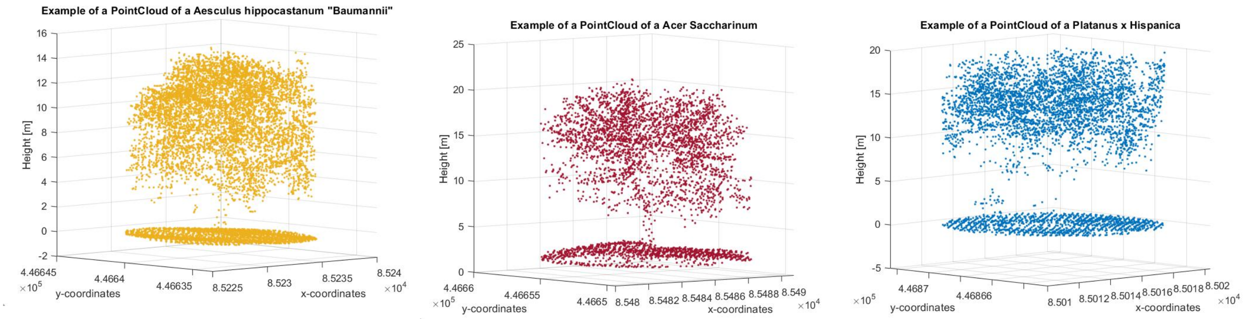


Figure 3.4: Buffer around multiple trees in Delft

The point cloud data that is included in the buffer, is subtracted and saved as an individual tree. In figure 3.5 are examples of different species shown. This plot is made with the *scatter3* function in *Matlab*. In figure 3.5a an example of the *Aesculus Hippocastanum* is shown. In figure 3.5b an example of the *Acer Saccharinum* and in figure 3.5c of the *Platanus x Hispanica* is presented.



(a) Point cloud of *Aesculus Hippocastanum*

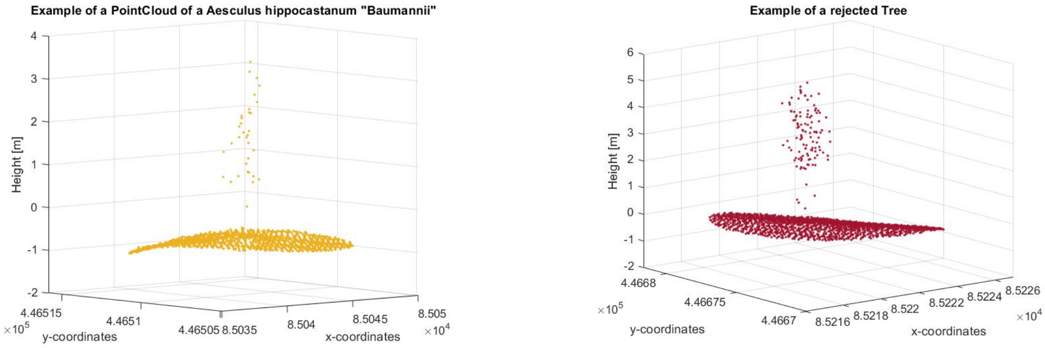
(b) Point cloud of *Acer Saccharinum*

(c) Point cloud of *Platanus x Hispanica*

Figure 3.5: Point cloud examples of individual trees, in Delft

3.3. Features of the Trees

In this research 6 different characteristics are estimated, whereby 4 of these characteristics have an normalized version. Each of this characteristics is described in the following paragraphs, but before the features can be calculated the point cloud data needs to be analyzed. In the previous step, the buffering of individual trees is done, but a selection of suitable trees should be made. Only fully developed and sufficiently sampled trees can be used for this research. To make this selection, the number of observations that are attributed to a tree must be examined, and which points are *tree* and which points belong to the *ground*. In figure 3.6, two examples of unsuitable trees are shown.



(a) Point cloud data from a rejected tree

(b) Point cloud data from a rejected tree

Figure 3.6: Point cloud data examples from non-approved trees

The first step in the process of separating the appropriate trees, is to remove all point cloud buffers with less than 5000 points. This quantity is based on a trial-and-error process, in which the result must contain sufficient points to be able to carry out further data analysis. [1], [7], [13]

The following step is to arrange the trees in different layers. The *ground level*, *upper canopy boundary* and the *lower canopy boundary* are the layers of interest. To determine these layers, per tree is a vertical histogram made. In figure 3.7 one can see an example for an *Aesculus Hippocastanum* tree. On the y-axis is the height of the bins, and on the x-axis is the amount of points that is counted for, in the corresponding bin. The histogram is made with 70 bins. The choice of this number is based on the relation between visualizing sufficient information about the heights and what shows a clear and structured result.

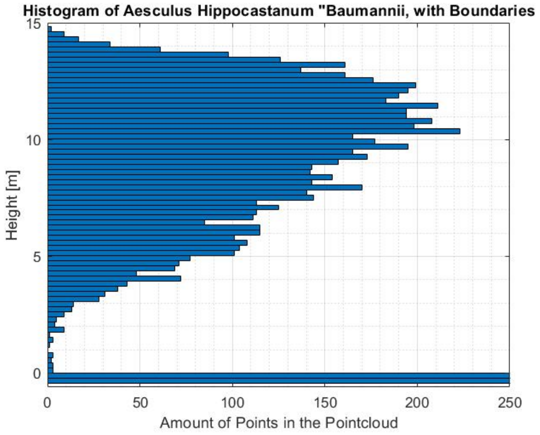


Figure 3.7: Example histogram of the *Aesculus Hippocastanum*

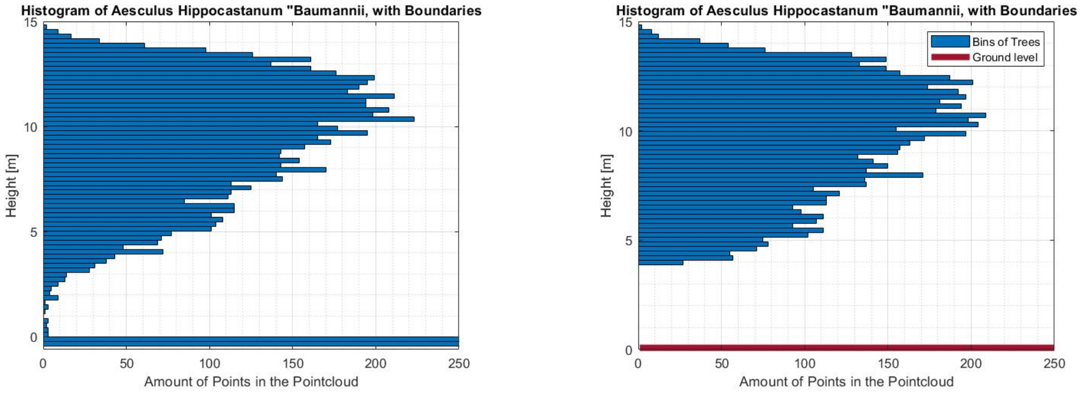
Such a histogram is made for all the trees that have at least 5000 measurement points. The next stage in pre-processing the individual point cloud data sets, is to determine the ground level.

Ground Level

The AHN point cloud data does not show one value for the ground, but shows a range of values that can be defined as *ground level*. The weighted mean is calculated for values that are under the height of 2 meters. This number is based on the fieldwork results. The weighted mean is calculated using the formula:

$$\overline{x} = \frac{\sum_{n=1}^{n}(x_n * w_n}{\sum_{n=1}^{n} w_n} \tag{3.1}$$

Whereby the w represents the weight of the value x. The weight is the amount of points that represent an certain height, in other words the corresponding bin value. The n is the amount of bins that are taken into account. The bins that correspond with the *ground layer* are replaced with this weighted mean. This results in the histogram shown in figure 3.8.



(a) Original Histogram

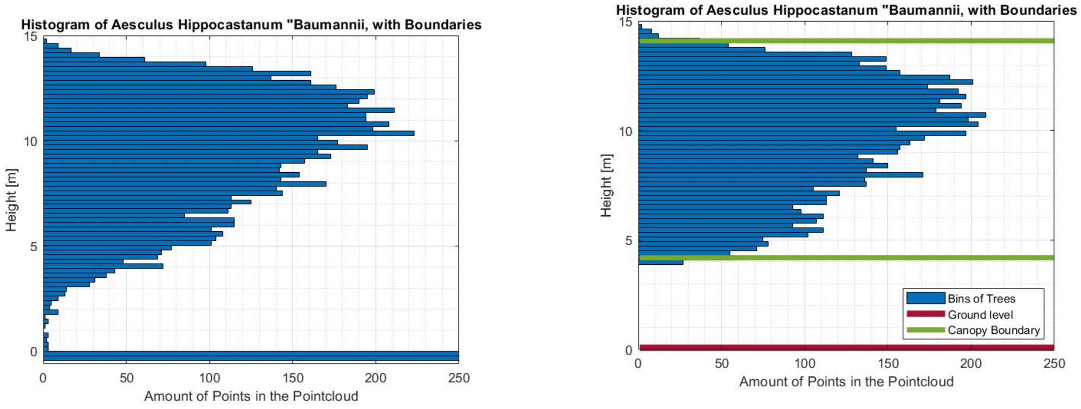
(b) Histogram with the determined Ground Level

Figure 3.8: Example Histogram of the *Aesculus Hippocastanum*

A point cloud is only treated as a tree if at least 2500 points above the ground have been measured. If a point cloud does not meet this requirement, it is removed. [1], [7], [13]

Upper- and Lower- Canopy Layer

For further analysis of the point cloud buffers, the canopy needs to estimated. For this estimation, the vertical histogram is used. If there are at least 22 points in one bin, the bin is considered to be canopy. The upper boundary is determined as the highest bin, and the lower boundary as the lowest bin with at least 22 measurements points. This quantity is based on a field study, whereby the height of the lower canopy boundary is measured for different trees. These measurements varied between 2,5 and 6 meter, which corresponded for almost all the trees with a value of 22 points in the bin.
In figure 3.9 the original histogram is shown, and on the right side (figure 3.9b) the histogram with the boundaries.



(a) Original Histogram

(b) Histogram with the determined Ground Level and Canopy Boundaries

Figure 3.9: Example Histogram of the *Aesculus Hippocastanum*

This ground level and the canopy boundaries are used to determine the different features.

Height of the tree

The characteristic *Height* of the tree is the most intuitive, and easiest to calculate. As described in the previous paragraph, the upper boundary of the canopy and the ground level are determined. The height of the tree is the difference between this upper boundary and the ground. In formula form this becomes:

$$\text{Height} = \text{Upper Canopy Boundary} - \text{Ground Level} \tag{3.2}$$

This is visualized in figure 3.10.

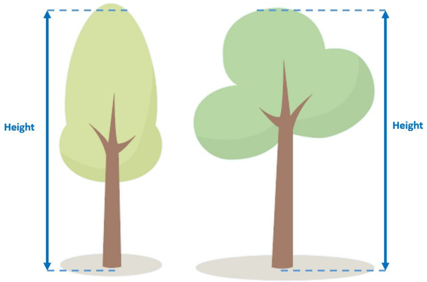


Figure 3.10: Height of the Tree

The height of the tree is dependent on the age. In general one can say that the tree becomes bigger, and higher by age. Therefore there is sometimes a correction needed, or a scaling. These "normalized" values are relevant for the next features.

Height of the Trunk

The *Trunk Height* is determined as the distance between the ground and the lower boundary level. In formula form this is:

$$\text{Trunk Height} = \text{Lower Canopy Boundary} - \text{Ground Level} \tag{3.3}$$

The distance between the canopy boundary and the ground is shown in figure 3.11

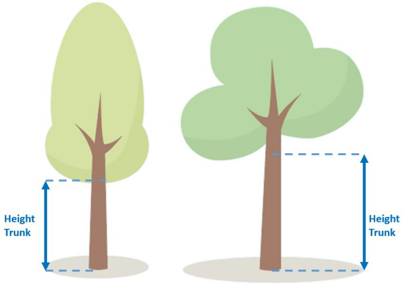


Figure 3.11: Height of the Trunk

To avoid the influence of the size of a tree on this parameter, the estimated parameter is normalized. This is calculated using the formula:

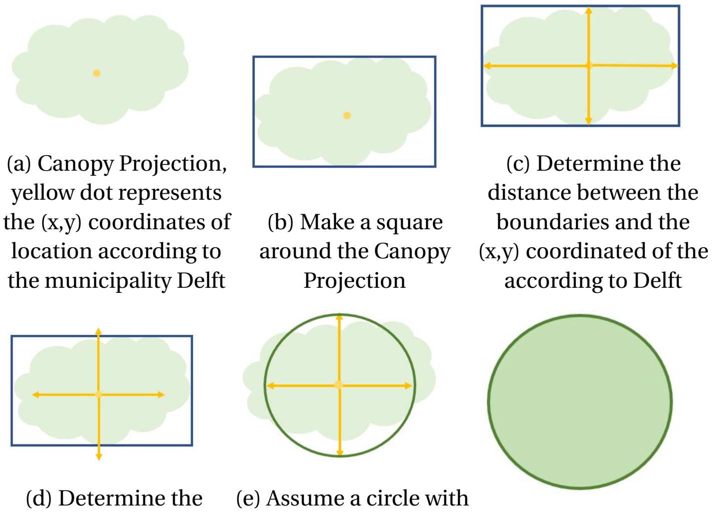
$$\text{Normalized Trunk Height} = \frac{\text{Trunk Height}}{\text{Height Tree}} \tag{3.4}$$

Canopy Projected Area

The *Canopy Projected Area*, further called CPA, is estimated by first projecting points onto a horizontal plane. Then, the boundary of the points is estimated by determining the mean radius. First the distance between the boundary of the square is calculated, where after the mean distance is calculated. This will be seen as the radius of a circle, whereby the area is calculated using the formula:

$$Area = \pi * \text{radius}^2 \tag{3.5}$$

In figure 3.12 the steps are described.



(a) Canopy Projection, yellow dot represents the (x,y) coordinates of location according to the municipality Delft

(b) Make a square around the Canopy Projection

(c) Determine the distance between the boundaries and the (x,y) coordinated of the according to Delft

(d) Determine the mean distance between the yellow dot and the boundaries

(e) Assume a circle with the radius the mean radius, determined in the previous step

(f) Determine the area according the given formula

Figure 3.12: Step wise explanation of the determination of the Canopy Projected Area

The resulting area contains information about the spreading of the canopy. This feature is shown in figure 3.13.

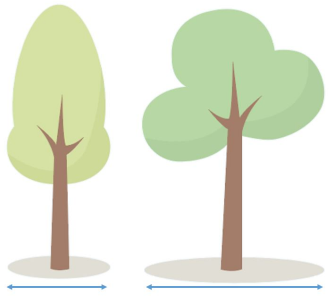


Figure 3.13: Canopy Projected Area

The CPA can be influenced by the height of the tree, therefore one needs to normalize this characteristic as well. This is done using the formula:

$$\text{Normalized CPA} = \frac{\text{CPA}}{\text{Height Tree}} \tag{3.6}$$

Ratio of Diameters

First the ratio between the diameters in the x- and y-direction are determined. The first three steps that are done to determine the CPA, are now used to determine the diameters. After the square outline is made (figure 3.12c), the difference in the x- and y-direction are estimated. The ratio of these diameters is computed by:

$$\text{Ratio of Diameters} = \frac{\text{Diameter in y-direction}}{\text{Diameter in x-direction}} \tag{3.7}$$

To avoid the influence of the size of a tree on this ratio, the estimated parameter is normalized as well. This is done according to:

$$\text{Normalized Ration} = \frac{\text{Ratio of Diameters}}{\text{Height of the Tree}} \tag{3.8}$$

Centre of Gravity distance

In the chapter *Background Information* the meaning of the *Centre of Gravity* is already explained. However, not the vector from the beginning of the trunk to the centre of gravity is of essence. The horizontal distance between the projection of the centre of gravity and the (x,y) coordinates of the trunk, according to municipality Delft is used instead. This characteristic is very sensitive to errors in the determination of the (x,y) coordinates by the municipality of Delft. More about this in chapter *Discussion*.

The first step in these calculations is to determine the centre of gravity, this is done by taking the mean of all the data points that are defined as tree. This results in one value, that is -almost- in the middle of the data set. The (x,y) coordinates of this value are taken and the distance to the trunk is calculated. This characteristic is not only related to the height of the tree, but especially to the direction of growth. In figure 3.14 one can see a simplification of the situation. The red arrow is the concerning distance.

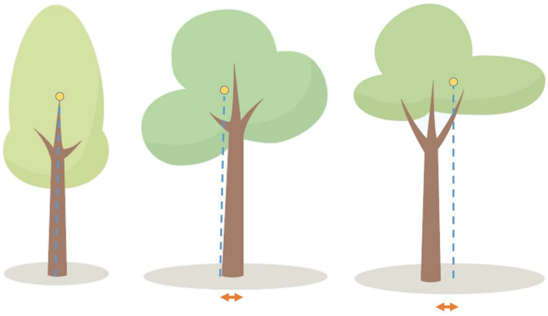


Figure 3.14: Centre of Gravity Distance. Arrow is the distance, yellow dot the centre of gravity.

This characteristic depends on the height of the tree, to correct for that the normalized value is taken into account. This is done using the formula:

$$\text{Normalized Centre of Gravity} = \frac{\text{Distance Centre of Gravity to Trunk}}{\text{Height Tree}} \tag{3.9}$$

This feature is further called "CGrav distance" or "Centre of Gravity distance".

These features are calculated for all the trees of interest in Delft. This results in one table, which is used in the classification process.

3.4. Classification of Tree Species

In the previous step a table with all the different features is created. For each thee, the following characteristics are known: *Height, Trunk Height, Normalized Trunk Height, CPA, Normalized CPA, Ratio of Diameters, Normalized Ratio of Diameters, Distance between Centre of Gravity Projection and the normalized value of that characteristic.* The following step in the process of classification is to apply the *Random Forest Classification.* Therefore a training- and validation- data set is distinguished from the matrix with features. The training data set exist of the features of 60 trees, per tree of interest. In total there are 180 trees in the training set. The validation data is set to be 30 trees, per tree of interest. In total 90 trees.

3.4.1. Random Forest Classification

Random Forest Classification is a method of supervised classification, based on decision trees. This algorithm creates not only one decision tree, but a forest with a number of trees. In general, the more trees in the forest, the more robust the forest is. Decision trees are an emerging method for various machine learning tasks. In these tree models, the actual shape of a tree can be found. The beginning of the trunk is the first decision to be made, the branches represents conjunctions of features that are lead to class labels, represented by the leaves. Each node in the tree corresponds to one of the input variables; it creates a binary split. The splitting process is done when certain criteria are met [12]. For the classification of trees in Delft, the following example can make it more understandable. One is walking around in Delft and sees a nice tree, and it is known that this can only be classified as *Aesculus Hippocastanum*, *Acer Saccharinum* or a *Platanus x Hispanica*. To determine which tree is seen, several questions are asked. These questions (the nodes) lead to different answers (leaves) via the branches. The first question can be the height of the trunk, if it is bigger than the value x, it is supposed to be a *Platanus x Hispanica*, otherwise its a *Aesculus Hippocastanum* or *Acer Saccharinum*. To discriminate the last two species, one can look at the CPA. If this is bigger than value y, the tree is supposed to be a *Aesculus Hippocastanum* and otherwise it is an *Acer Saccharinum*. This is a very simplified decision tree, whereby the values of x and y are automatically generated by the decision algorithm. In figure 3.15, an simple representation of decision trees is shown. In the first tree, the outcome is *green*, as well as in the third tree. However, the outcome in the second tree is *orange*. The bias of these decision trees is quite big.

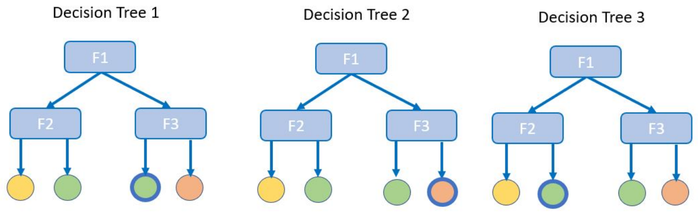


Figure 3.15: Example of Decision Trees, resulting in prediction of the *green* and *orange* species

One of the limitations, of classification based on decision trees, is the high variance. This variance can be reduced by the use of Random Forest Classification. Before the Random Forest Classification can be applied, the process of bagging (or bootstrap aggregating) is necessary. This bagging is actually the creation of a "forest", whereby several decision trees are combined and merged. The variance of the outcomes is reduced by taking the average of the results. The downside of bagging is that the risk occurs of correlation between different decision trees, increasing the bias in the model. The algorithm behind the Random Forest Classification reduces this correlation between decision trees, by choosing only a sample of the features at each split (node in the decision tree). This makes the trees no longer correlated, and the variance will reduce. [12]

For Random Forest Classification, multiple decision trees are averaged. In the case of classification of different species of trees, the outcome of the decision tree can only be the name of a species. Therefore the most common outcome is the determined species [12]. In figure 3.16 one can see a summary of the random forest classification. There are only three decision trees taken into account, but the conclusion is the *green* species.

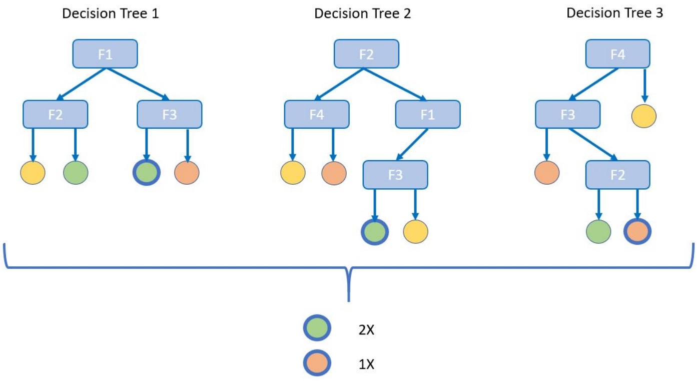


Figure 3.16: Example of Random Forest Classification, resulting in prediction of the *green* species

Implementation of the Random Forest Classifier

As told before, the matrix with all the features is split into an training data set, called *TrainingFeatures*, and the validation set, called *ValidationFeatures*. This is done for the species of the trees as well; *TrainingSpecies*, and *ValidationSpecies*. The implementation is done in *Matlab* by the *treebagger* command. The number of decision trees is set to be 40. The choice to use 40 decision trees, so that the variation in the resulting classification is minimized. A large amount of decision trees leads to a better result, until correlation between decision trees occurs. For 40 decision trees, the correlation does not appear and the amount is still understandable for the human mind. One should take into account that there are only 5 features, thus the combinations and the sequence of nodes is not limitless.

The prediction is made by the *predict* commend, where the decision tree and the validation features are taken into account. In the chapter *Results*, the results of the Random Forest Classification are presented.

3.4.2. Confusion Matrix

The confusion matrix is a performance measurement for classification problems, where the output can be two or more classes. It is a table with -at least- four different combinations of predicted and true values. The table shows 4 different cells, with each a *True Positive, True Negative, False Positive and True negative*. Whereby the *True Positive* the interpretation of the decision tree is that the prediction is 'true' and it is actually true. The *True Negative* is that the interpretation of the decision tree is that is it negative, and that is actually the case. The *False Positive* also called *Type 1 Error* is that the prediction is positive, but the actual outcome should be negative. For the *False Negative* or *Type 2 Error*, the prediction is negative, but it is actually positive. However, in the case of this research there are not two, but three different classes. This makes things more complicated, but still manageable. In figure 3.17 one can see an example of a confusion matrix, with the three different species it in.

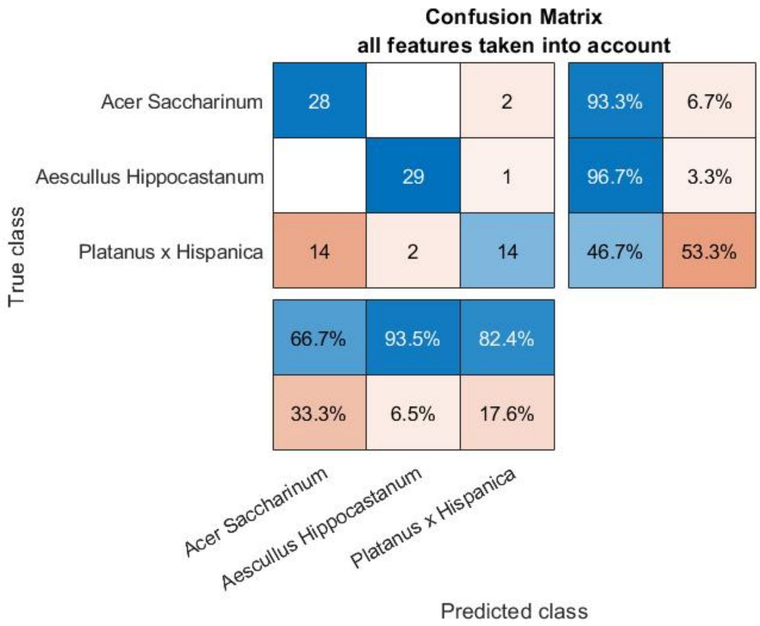


Figure 3.17: Example of Confusion Matrix in this research

A short explanation: the Random Forest Classification predicts the point cloud data to be an *Acer Saccharinum*, and in 28 times this is true. And in 14 times it is actually an *Platanus x Hispanica*. And the Random Forest Classifier predicts 2 times that the tree is an *Platanus x Hispanica*, but that should actually be an *Acer Saccharinum*. On the right and bottom side, one can see (in blue) how much percent of the time the true class is predicted. And how many times it got wrong (orange).

Implementation of the Confusion Matrix

The confusion matrix is easy to implement in *Matlab*, with the command *confusion matrix(Predicted Values, Training Values)*. The results of the implementations are shown in the chapter *Results*.

3.4.3. Importance of Features

In the process of classification, five features have been used to far. However, an analysis of the effect of these features is required. This section explains how the effects can be determined.

Height
The height of the tree depends on the age, and is therefore not suitable for classification. Suppose the *Platanus x Hispanica* does not exceed 15 meters and a *Acer Saccharinum* does. If the point cloud of a random tree shows a height of 15 meters, it is indistinguishable whether it is a mature *Platanus* or a younger *Aesculus*. Therefore the characteristic *Height* is not taken into account. The leftover features are: *Normalized Trunk Height, Normalized CPA, Normalized Ratio of Diameters,* and *Normalized Centre of Gravity Distance*.

Further study of the importance features is done by analyzing the decision trees, generated for the Random Forest Classification. This results in a bar graph, where the three most important features are used for classification based on clustering.

3.4.4. Clustering

Clustering is a type of unsupervised classification, which is used when the data is unlabeled. The general goal of this principle is grouping a set of objects in such a way that objects in the same group (the cluster) are more similar to each other than to the objects in other clusters or to objects that are not part of cluster. The results of clustering are very intuitive and easy to understand [2] [3] In this research, the three most important features (determined with analysis of the treebagger) are plotted against each other and clusters are formed. The results of this unsupervised classification are shown in the chapter *Results*.

4

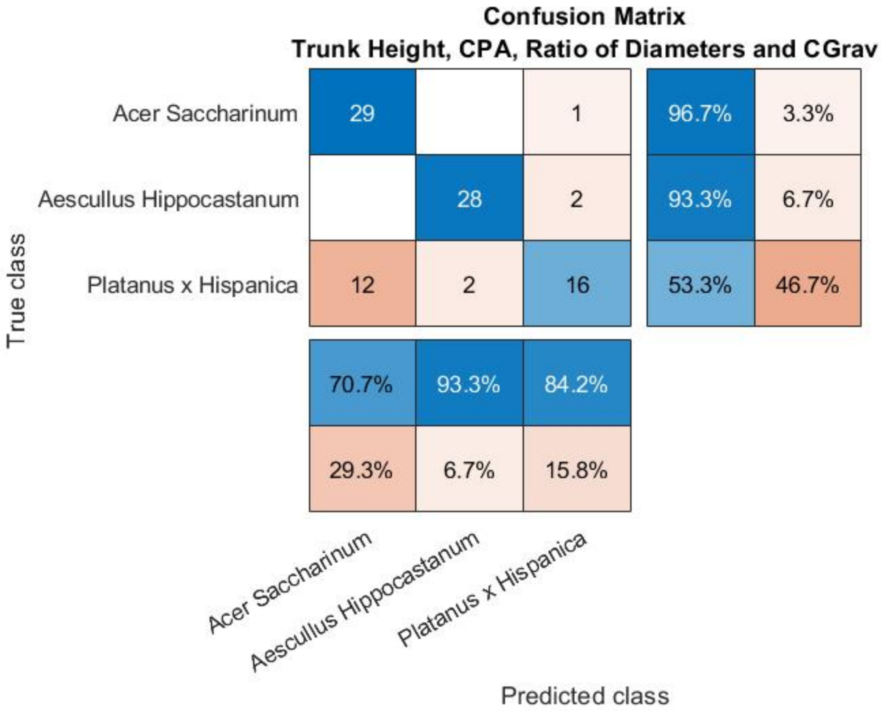
Results

The results of the supervised classification *Random Forest Classification* are presented in this chapter. First the results of the classification with four features taken into account are presented and afterwards with three different features. At last, the clustering classification is presented.

4.1. Results of Random Forest Classification

The method of Random Forest Classification is explained in the chapter *Methodology*. Initially the four most important features are used for classification, and a confusion matrix is calculated. In this chapter the confusion matrix is shown and discussed. In Appendix A, one can see an example of one of the corresponding decision trees.

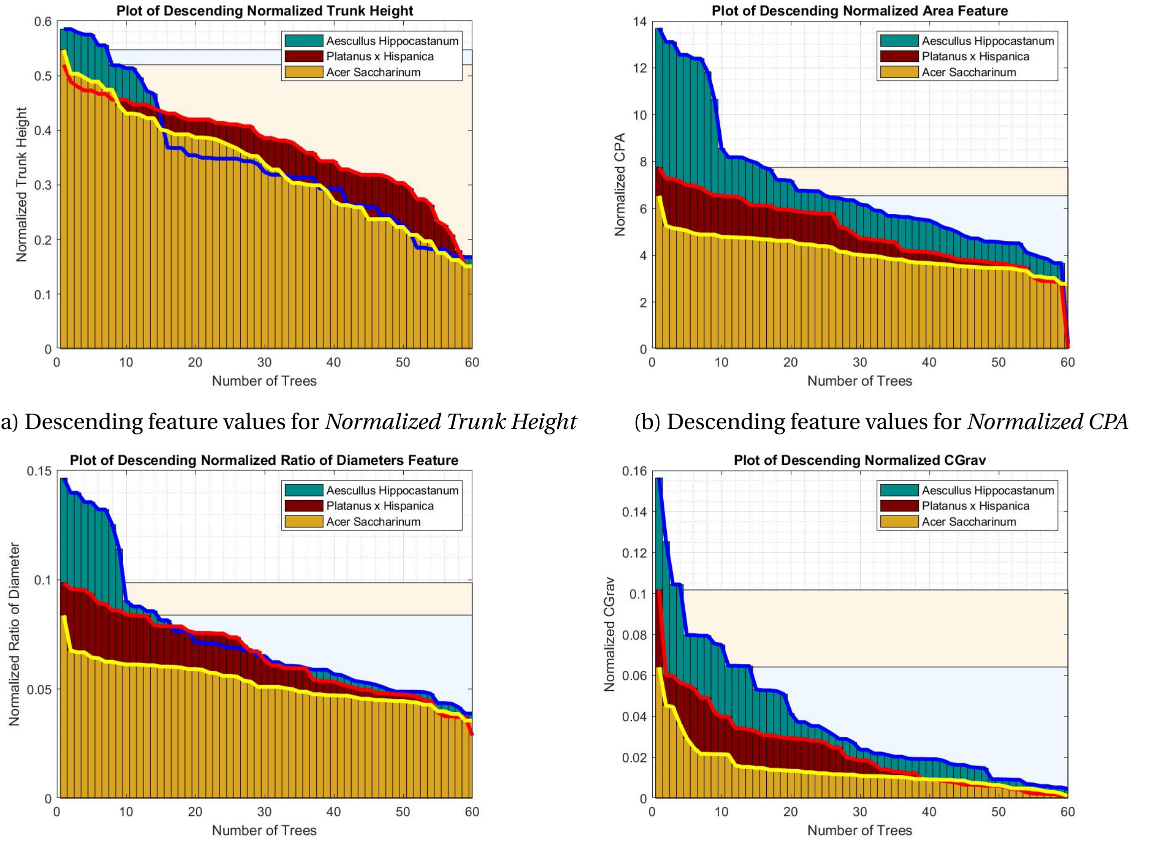
In figure 4.1 the confusion matrix is shown. Hereby the features *Normalized Trunk Height, Normalized CPA, Normalized Ratio of Diameters* and *Normalized Centre of Gravity Distance* are taken into account.



	Acer Saccharinum	Aescullus Hippocastanum	Platanus x Hispanica		
Acer Saccharinum	29		1	96.7%	3.3%
Aescullus Hippocastanum		28	2	93.3%	6.7%
Platanus x Hispanica	12	2	16	53.3%	46.7%
	70.7%	93.3%	84.2%		
	29.3%	6.7%	15.8%		

Figure 4.1: Confusion matrix

As can be seen, the *Aesculus Hippocastanum* is more than 90 percent correctly classified. And the classification for *Acer Saccharinum* is even more than 95 percent correct. The features that belong to these trees are very distinctive. The features of the *Platanus x Hispanica* are less characteristic. This can be displayed with the aid of graphs. In figure 4.2 these features are displaced. On the x-axis the amount of trees in the training data is shown on the x-axis. And on the y-axis the values of the feature of interest. The trees are sorted in descending order and displayed per species. The light blue and light pink areas in the background of the graph represent the area in which the feature is not distinctive. In the blue area the tree can be *Aesculus Hippocastanum* or *Acer Saccharinum*. In the pink area the tree can be *Aesculus Hippocastanum*, *Acer Saccharinum* or *Platanus x Hispanica*.



(a) Descending feature values for *Normalized Trunk Height*

(b) Descending feature values for *Normalized CPA*

(c) Descending feature values for *Normalized Ratio of Diameters*

(d) Descending feature values for *Normalized Centre of Gravity Distance*

Figure 4.2: Distinctive features for the species of interest

In figure 4.2a the Normalized Trunk Height is plotted. The *Aesculus Hippocastanum* can be distinguished for the highest values, but for a normalized trunk height of 0.05 lower, it becomes confusing with the *Acer Saccharinum*. And for values 0.1 lower it becomes confusing with both the *Acer Saccharinum* and the *Platanus x Hispanica*. That is why the blue and pink areas in the background of the graph show a lot of overlap. The values of the three different species are close to each other across the entire width of the graph and show a similar linear descending line.

Figure 4.2b shows the Normalized CPA plotted against the number of trees, in descending order. It is noticeable here that the *Aesculus Hippocastanum* is again the easiest to distinguish, it has unique values for the trees with the highest values for Normalized CPA. The values of the Normalized CPA of the *Acer Saccharinum* and *Platanus x Hispanica* are clearly distinguishable for about half of the trees. It is remarkable that the values of the Normalized CPA for all species go to about the same value. Only the smallest value of the *Aesculus* and *Acer* gives a different value, more about this in the chapter *Discussion*.

In figure 4.2c the values of the Normalized Ratio of Diameters is shown. Again the *Aesculus Hippocastanum* has the highest values and the *Acer Saccharinum* the lowest. The *Platanus x Hispanica* is in between.

The last figure, figure 4.2d shows the values of the Normalized Centre of Gravity Distance. The values are again differentiating for the *Aesculus Hippocastanum* and the *Acer Saccharinum*. The pink and blue areas in the background of the graph are further apart, but this is due to the largest value of the Platanus. This value is slightly different and is discussed further in the chapter *Discussion*.
Not every feature contributes as much to the classification process. To display this contribution, a "weight" is attached to the feature and figure 4.3 shows how important this feature is in the process. The features are represented by *x1, x2, x3 and x4* and stand for *Normalized Trunk Height, Normalized CPA, Normalized Ratio of Diameters and Normalized Centre of Gravity Distance* respectively.

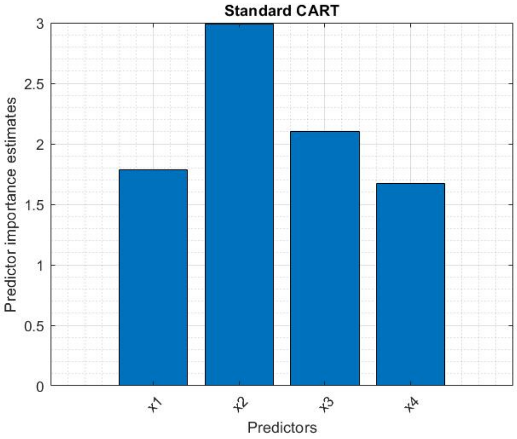


Figure 4.3: Importance of predictors

This shows that he Normalized CPA feature has the greatest influence on the classification of trees. Followed by the Normalized Ratio of Diameters and the Normalized Trunk Height. However the difference between Normalized Trunk Height and Normalized Centre of Gravity Distance is quite small. With these three features, Random Forest Classification has been implemented, where again a confusion matrix is calculated. This confusion matrix is shown in figure 4.4.

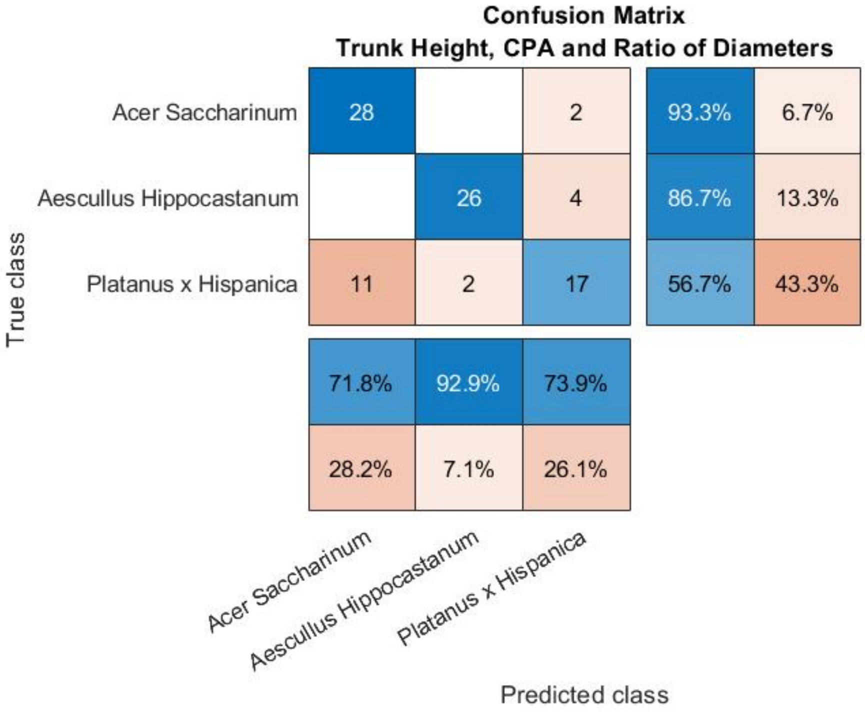


Figure 4.4: Confusion matrix with the three most important features

The differences between these two confusion matrices, and an example of the wrongly classified trees, is discussed in the chapter *Conclusion*.

4.2. Results of Clustering Classification

The following step is to find out whether these three features are sufficiently distinguished to classify the three species based on clusters. Therefore the properties are plotted against each other, and the different species are shown in different colours. The results of the clustering is shown in figure 4.5

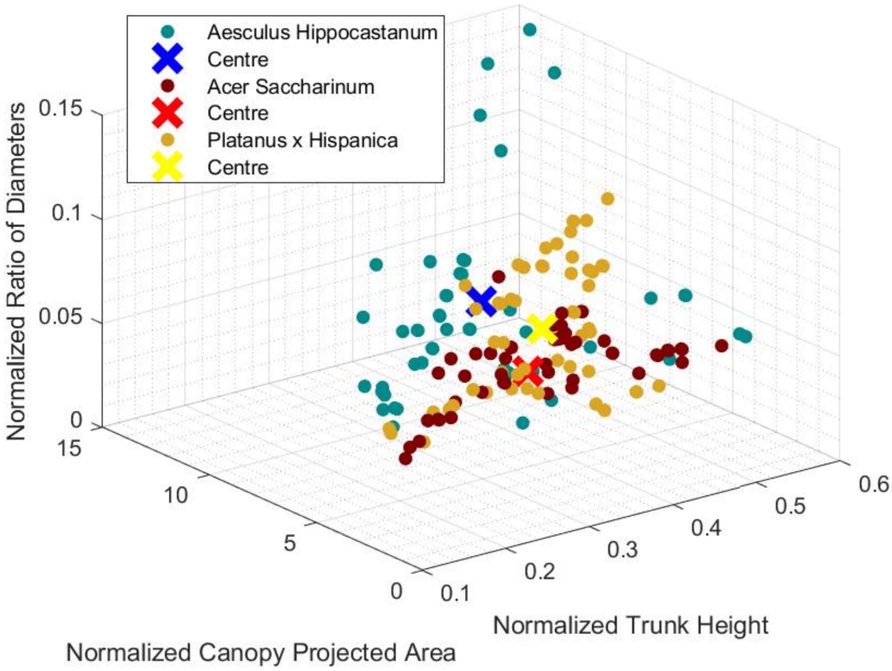


Figure 4.5: Clustering with Normalized Trunkheight, Normalized CPA en Normalized Ratio of Diameters taken into account.

In Appendix B the same clustering is shown, but from other angles.

5 Conclusion

From the results in the previous chapter, the conclusions and recommendations are made. The main research question is: **Is it possible to classify the "*Aesculus Hippocastanum*","*Acer Saccharinum*" and "*Platanus x Hispanica*", based on AHN3 data, in Delft?**.

The "*Aesculus Hippocastanum*","*Acer Saccharinum*" and "*Platanus x Hispanica*" are all found in Delft, on several locations. They can be identified manually by looking at different characteristics as the shape of the leaves and the colours of the trunk. This is a manually intensive process, that probably can be done more efficiently. To overcome this problem, AHN data is examined and features are determined based on this point cloud data set. Classification is done based on these extracted features.

5.1. Possibility to determine features in AHN data

Is it possible to determine different features of trees, based on AHN data?
The AHN point cloud data set provides enough information for determination of different features of trees. This research is an extension of a previous research [19], whereby mobile laser scanners were used. The characteristics determined based on this mobile laser scanner data set where not possible to determine by AHN data. Therefore some new features needed to be determined.

5.2. Characteristics of the species of interest

Which features of trees can be used to classify these trees of interest?
The features to discriminate these trees of interest are the height, *Normalized Trunk Height*, Normalized Canopy Projected Area, *Normalized Ratio of Diameters* and the normalized distance between the Centre of Gravity projection on the ground and the trunk. How these features are determined is described in the chapter *Methodology*. Using these features, classification can be applied. Since the height of the tree strongly depends on the age, this feature is not included in the classification process. Analysis of the remaining features shows that the *Normalized Trunk Height, Normalized CPA* en *Normalized Ratio of Diameters* are the most important predictors. As a final result, Random Forest Classification and Clustering have been implemented with these three features.

5.3. Supervised Random Forest Classification

Can classification of the three species of trees be done using the supervised Random Forest Classification?
Yes, it is possible to classify the three species of trees. The result is shown in figure 4.4, this result is based on classification done with three different features (Normalized Trunk Height, Normalized CPA en Normalized Ratio of Diameters). The result shows that the *Acer Saccharinum* and the Aesculus Hippocastanum can be distinguished form each other very well. The *Platanus x Hispanica* is more likely to go wrong. When looking at the graphs shown in figure 4.2 it appears that the *Platanus x Hispanica* has the most overlap between the trees. More about this classification result in the last section of this chapter.

5.4. Clustering

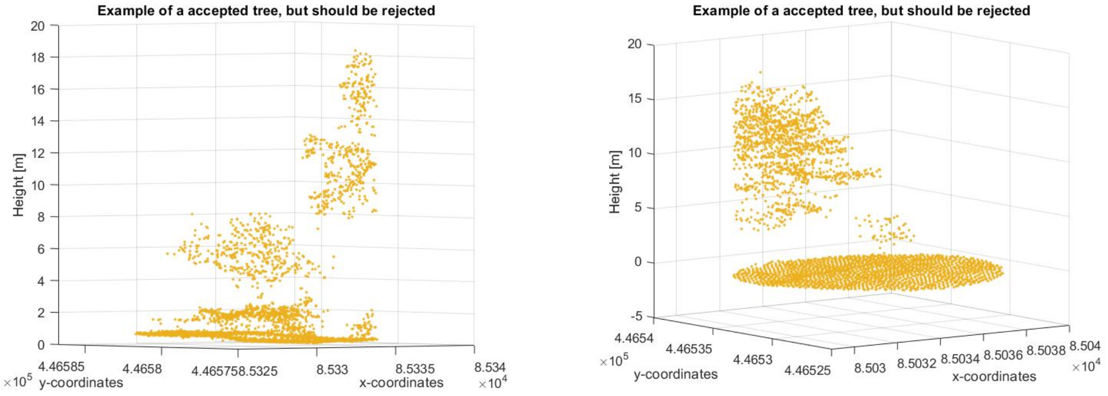
Is it possible to classify the three different species of trees, with clustering?

Classifying with clusters yield an unsatisfactory result. The different species cannot be distinguished from each other. In figure 4.5 the result of the clustering can be viewed. The centres of the clusters (the crosses in the figure) are too close to each other and the clusters overlap too much.

5.5. Wrongly classified trees

How do some of the wrongly classified trees look like? And why are they wrongly classified?

In figure 4.4 the results of the Random Forest Classification is shown. This shows that 21 out of the 90 trees are classified incorrectly. There are many causes of this incorrect classification; among other things the features may be not sufficiently distinctive, the AHN data is incorrect, the buffer of the point cloud is too big or too small, or there is an obstacle what interferes with the tree. Another possibility is that the tree meets the requirements of a tree, but is actually not a tree. Research shows that the point cloud data of the wrongly classified trees often do not represent a tree. In figure 5.1 two examples are shown. These trees meet the conditions of the amount of points in the point cloud and the amount of points in the canopy. However, if the form is viewed, it does not appear to be healthy developed trees.



(a) Treated as tree, but it should be rejected as tree

(b) Treated as tree, but it should be rejected as tree

Figure 5.1: Examples of wrongly classified trees

A small field study shows that at the location indicated by the municipality of Delft, no tree appears to be standing at these locations. Here over more in the chapter *Discussion*.

5.6. Final Conclusion

Is it possible to identify the "*Aesculus Hippocastanum*","*Acer Saccharinum*" and "*Platanus x Hispanica*", based on AHN3 data, in Delft?

Yes, is it is possible to classify the "*Aesculus Hippocastanum*", "*Acer Saccharinum*" and "*Platanus x Hispanica*" based on AHN3 data, in Delft. The *Aesculus Hippocastanum* is with 86.7 percent accuracy to classify, the *Acer Saccharinum* with 93.3 percent and the *Platanus x Hispanica* with 50 percent. This classification is done with the Random Forest Classification method, using the features; *Normalized Trunk Height*, *Normalized CPA*, and *Normalized Ratio of Diameters*. These three different species can not be sufficiently classified using clustering.

6 Discussion and Recommendations

In this discussion the different finding of this research are reviewed and discussed.

6.1. Debatable measurements in the provided data sets

During this research two different data sets were used. The one obtained from the municipality of Delft, with the location of the trees, may contain ambiguities. The location of the tree is indicated by a point, in a very precise x- and y-coordinate, while the trunk covers an area. It is unknown how exactly the location is determined, and whether it is done in a consensual way. This lack of clarity mainly affects the determination of the *Centre of Gravity Distance* feature. Hereby the precise distance between the projection of the centre of gravity on the horizontal plane and the provided (x,y)-coordinates is calculated. If there are inconsistencies or errors in the location of the municipality, this has far-reaching consequences for the distance concerned. For this particular research it does not matter which point from the trunk was taken, as long it's consequently the same point.
Another disadvantage of using this database is that it is unknown when and how often it is updated. It may be that something happens to the tree between the time the municipality takes measurements and the AHN data is collected. But also in the other order: that something happened to the tree between collecting the AHN data and updating the database. Some examples of such events are the fall of the tree (by storm or parasites), lighting strike or by cutting down the tree. Better results can be obtained by a more extensive field research. There can then be focused on finding the exact location and measuring it in a consistent way.

The other data set used is the AHN point cloud. Before it it obtained are all the points in the point cloud classified and labeled. For this research only points with the label "vegetation" were looked at. However, no verification has been made weather these points actually belong to vegetation and therefore is these points can belong to a tree. Only if the classification of points often goes wrong, it can be a problem for the kind of research that has been carried out. Keep in mind that at least 5000 points belong to 1 tree, otherwise it is not treated as a tree at all.
A bigger problem with the AHN data set is that all data is obtained from above. The elevation data is collected by flying over the land with an airplane or helicopter. The upper part of the tree is therefore best measured. Only when the laser comes through between the branches and the leaves, it can reach lower parts in the tree. This makes the collected point cloud "top heavy". The height of the tree can therefore be determined, but the determination of the lower canopy can therefore cause problems. Especially because the trunk is almost impossible to measure from above. An analysis of the point clouds (after buffering) and the vertical histogram indicates that this is not a very big problem. The shape of the tree is still clearly recognizable. In the research of Jinhu Wang and Roderik Lindenbergh, a feature was used that is related to the width of the trunk. This feature could be determined by mobile laser scanning, but with the use of AHN this is impossible.

6.2. Assumptions made

Several assumptions have been made during this investigation. The most influential assumptions are related to how many points there should be in a tree and how many points in the canopy. In this study there should be at least 5000 point in a buffer and at least 2500 points in the canopy. These values are based on a field study and a trial-and-error process. It was manually checked if the point cloud had the shape of a tree and if there was enough data to carry out a responsible data analysis. These numbers can be chosen more precisely and may differ per species. This could be an important part of a subsequent study.

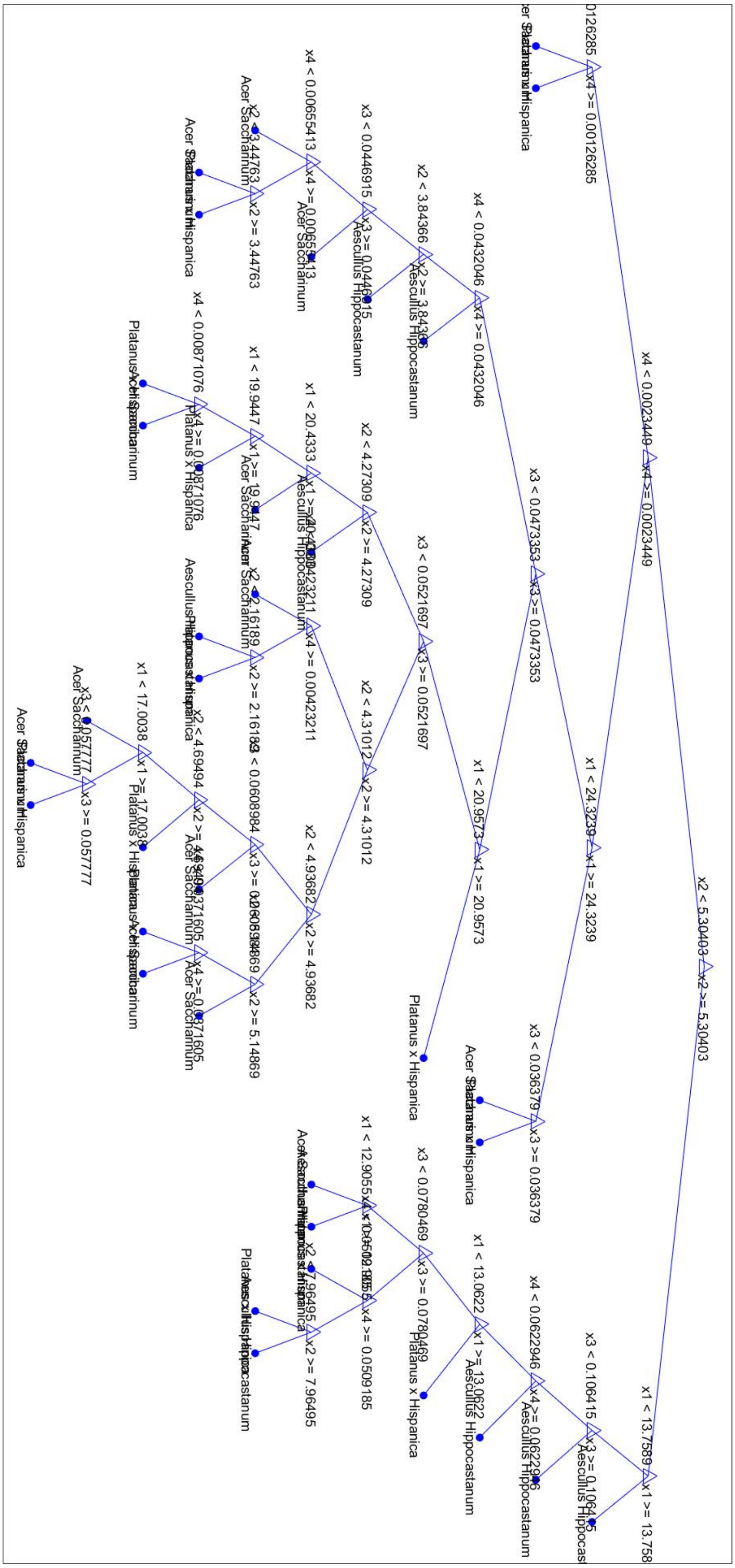
As can be seen in figure 3.4, a buffer has been made per tree. If the trees are close to each other, it is possible that the buffers will overlap. It is a possibility that an observation is included in the buffer of one tree, but actually belong to the other tree. The trees are so close together that only the points in the furthermost part of the canopy can be confused. The calculations of the features *Height* and *Trunk Height*, is done with the help of the *upper canopy boundary* and the *lower canopy boundary*. In determine these features, it wrongly including a point in a buffer is no big problem. However for the *CPA*, *Ratio of Diameters* and the *Centre of Gravity Distance* it is a problem. In a future study it could be examined whether it is possible to use a different radius for each buffer. Thus every buffer for a tree has another radius. Or a whole new method of indicating individual trees can be developed.

6.3. Further recommendations

In line with previous studies, it was decided to try to classify the *Aesculus Hippocastanum*, *Acer Saccharinum* and the *Platanus x Hispanica*. Another interesting step could be to choose three other trees. And afterwards to choose more than three trees and try to distinguish them form each other. Another angle for future research could be to focus on trees outside of Delft. It would be excellent if -in the end- the classification process is independent of the location and many trees can be distinguished at the same time.
Calculating the features is a time-consuming process. This could be accelerated if faster computers could be used. The University of Technology in Delft has such computers at its disposal, but they can not often be used for an additional thesis. But the early submission of a request should make it possible.

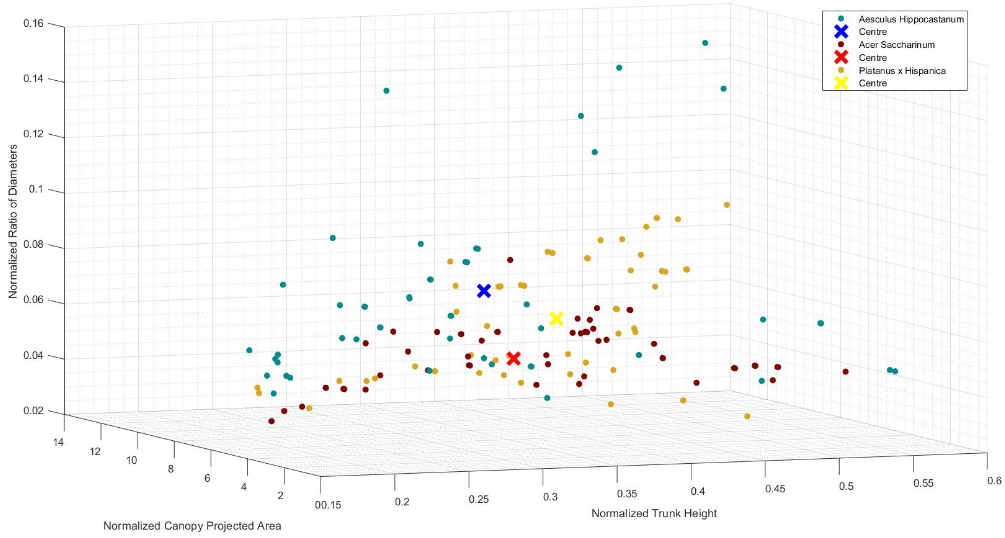
A

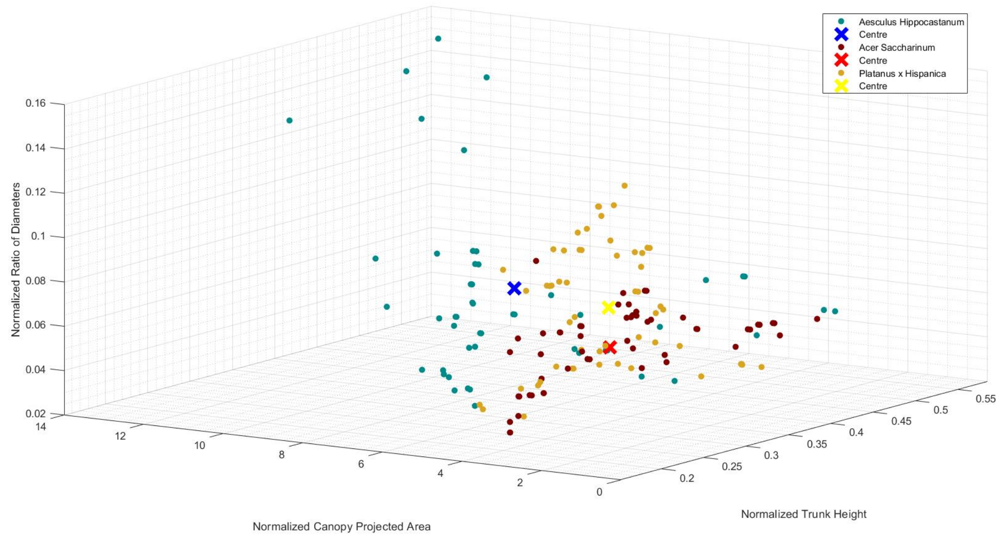
Example of a Decision Tree

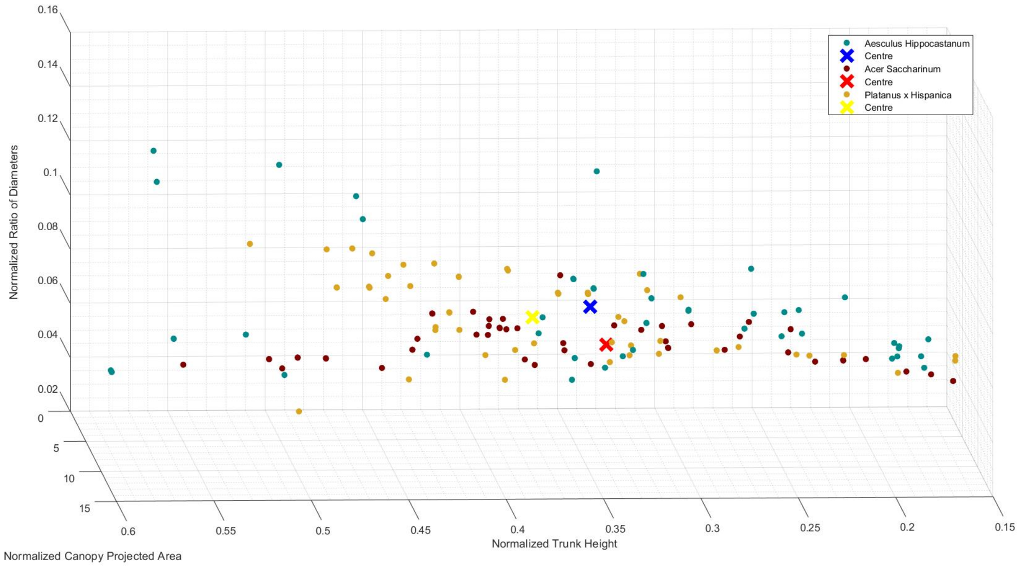

x2 < 5.30403
x2 >= 5.30403
x4 < 0.0023449
x4 >= 0.0023449
x1 < 13.7589
x1 >= 13.758
x4 >= 0.0012685
Platanus x Hispanica
x4 < 0.0432046
x4 >= 0.0432046
x3 < 0.0473353
x3 >= 0.0473353
x2 < 3.84366
x2 >= 3.84366
Aesculus Hippocastanum
x3 < 0.0446915
x3 >= 0.0446915
x4 < 0.00655413
x4 >= 0.00655413
Acer Saccharinum
x2 < 3.44763
x2 >= 3.44763
x3 < 0.0521697
x3 >= 0.0521697
x2 < 4.27309
x2 >= 4.27309
x1 < 20.4333
x1 < 19.9447
x1 >= 19.9447
x4 < 0.00871076
x4 >= 0.00871076
x4 >= 0.00423211
x2 < 2.16189
x2 >= 2.16189
x2 < 4.31012
x2 >= 4.31012
x2 < 4.93682
x2 >= 4.93682
x3 < 0.0608984
x2 < 4.69494
x1 < 17.0038
x1 >= 17.0038
x3 < 0.057777
x3 >= 0.057777
x2 >= 5.14869
x1 < 20.9573
x1 >= 20.9573
x1 < 24.3239
x1 >= 24.3239
x3 < 0.036379
x3 >= 0.036379
x3 < 0.106415
x3 >= 0.106415
x4 < 0.0622946
x4 >= 0.0622946
x1 < 13.0622
x1 >= 13.0622
x3 < 0.0780469
x3 >= 0.0780469
x1 < 12.9055
x4 >= 0.0509185
x2 >= 7.96495
Acer Saccharinum
Aesculus Hippocastanum
Platanus x Hispanica

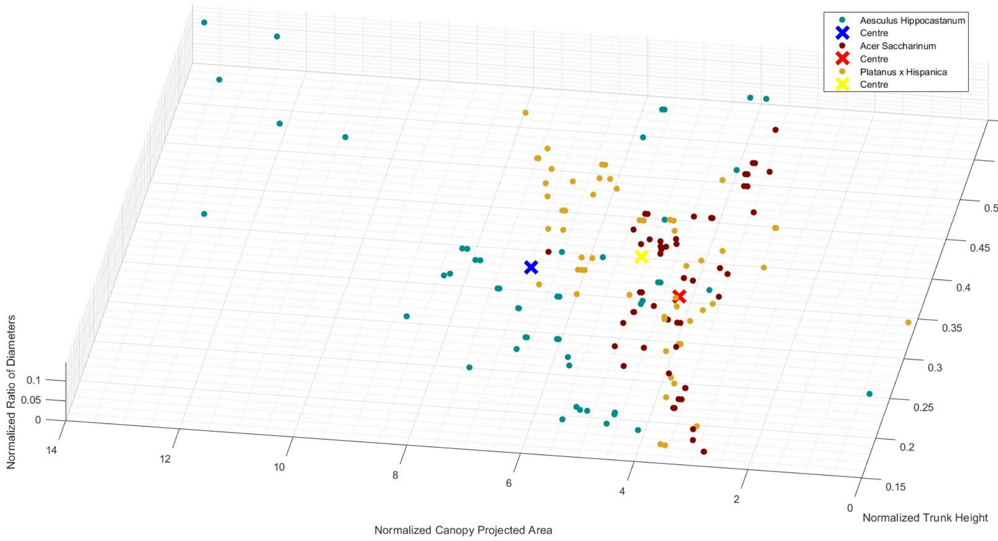
B

Clusters


Aesculus Hippocastanum
Centre
Acer Saccharinum
Centre
Platanus x Hispanica
Centre
Normalized Ratio of Diameters
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
14
12
10
8
6
4
2
00.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55
0.6
Normalized Canopy Projected Area
Normalized Trunk Height


Aesculus Hippocastanum
Centre
Acer Saccharinum
Centre
Platanus x Hispanica
Centre
Normalized Ratio of Diameters
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
14
12
10
8
6
4
2
0
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55
Normalized Canopy Projected Area
Normalized Trunk Height


Aesculus Hippocastanum
Centre
Acer Saccharinum
Centre
Platanus x Hispanica
Centre
Normalized Ratio of Diameters
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
5
10
15
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
Normalized Trunk Height
Normalized Canopy Projected Area


Aesculus Hippocastanum
Centre
Acer Saccharinum
Centre
Platanus x Hispanica
Centre
Normalized Ratio of Diameters
0.1
0.05
0
14
12
10
8
6
4
2
0
Normalized Canopy Projected Area
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
Normalized Trunk Height

List of Figures

Bibliography

[1] Bomenbieb - zilveresdoorn - acer saccharinum: informatie en foto's van de boomsoort zilveresdoorn op bomenbieb.nl. URL http://www.bomenbieb.nl/boomsoort/zilveresdoorn-acer-saccharinum/.

[2] An efficient data clustering method for very large databases, volume = 2, pages = 6, author = Tian Zhang, Raghy Ramakrishnan, Miron Livny, date = 2002, langid = english. .

[3] Learning with hypergraphs: Clustering, classification, and embedding. page 8, .

[4] Datasets - PDOK. URL https://www.pdok.nl/datasets.

[5] LAStools: converting, filtering, viewing, processing, and compressing LIDAR data in LAS format. URL http://www.cs.unc.edu/~isenburg/lastools/.

[6] T. van Beuningen H. van Meijeren A.K. Bregt, L.Grus. Wat zijn de effecten van een open actuaal hoogtebestand nederland. page 53.

[7] De Tuinen van Appeltern. Aesculus hippocastanum 'baumannii' (dubbelbloemige paardenkastanje). URL https://appeltern.nl/nl/shop/groen/tuinplanten/bomen/aesculus_hippocastanum_baumannii_dubbelbloemige_paardenkastanje.

[8] Faure Bernard Brenière Yvon Caron, Olivier. Estimating the centre of gravity of the body on the basis of the centre of pressure in standing posture. 30(11):1169–1171. ISSN 00219290. doi: 10.1016/S0021-9290(97)00094-8. URL http://linkinghub.elsevier.com/retrieve/pii/S0021929097000948.

[9] Thomas Gschwantner, Klemens Schadauer, Claude Vidal, Adrian Lanz, Erkki Tomppo, Lucio di Cosmo, Nicolas Robert, Daisy Englert Duursma, and Mark Lawrence. Common tree definitions for national forest inventories in europe. 43(2). ISSN 22424075. doi: 10.14214/sf.463. URL http://www.silvafennica.fi/article/463.

[10] S. Jennings. Assessing forest canopies and understorey illumination: canopy closure, canopy cover and other measures. 72(1):59–74. ISSN 0015-752X, 1464-3626. doi: 10.1093/forestry/72.1.59.

[11] E.T. Kanemasu. Seasonal canopy reflectance patterns of wheat, sorghum, and soybean. 3(1):43–47. ISSN 00344257. doi: 10.1016/0034-4257(74)90037-6. URL http://linkinghub.elsevier.com/retrieve/pii/0034425774900376.

[12] Wiener Matthew Liaw Andy. Classication and regression by randomForest. 2:6.

[13] Gregg McIntosh, Miriam Gómez-Paccard, and María Luisa Osete. The magnetic properties of particles deposited on platanus x hispanica leaves in madrid, spain, and their temporal and spatial variations. 382(1):135–146. ISSN 00489697. doi: 10.1016/j.scitotenv.2007.03.020. URL http://linkinghub.elsevier.com/retrieve/pii/S004896970700366X.

[14] Ross F. Nelson Erik Hans Ole Ørka Nicolas C. Coops Thomas Hilder Christopher W. Bater Terje Gobakken Michael A. Wulder, Joanne C. Wite. Lidar sampling for large-area forest characterization: A review. 121: 196–209. ISSN 00344257. doi: 10.1016/j.rse.2012.02.001. URL http://linkinghub.elsevier.com/retrieve/pii/S0034425712000855.

[15] Sorin C. Popescu. Estimating biomass of individual pine trees using airborne lidar. 31(9):646–655. ISSN 09619534. doi: 10.1016/j.biombioe.2007.06.022. URL http://linkinghub.elsevier.com/retrieve/pii/S0961953407001316.

[16] Kurt S. Pregitzer. Fine roots of trees - a new perspective. 154(2):267–270. ISSN 0028-646X, 1469-8137. doi: 10.1046/j.1469-8137.2002.00413_1.x. URL http://doi.wiley.com/10.1046/j.1469-8137.2002.00413_1.x.

[17] Kirsten van Dongen, Merve Sinem Gunes, and Esther Roosenbrand. CIE4614 3d surveying of civil and oshore infrastructure big assignment AHN tree analysis. page 19.

[18] Christine Azevedo-Costo Mitsuhiro Hayashibe Sébastien Cotton Phillipe Fraisse Vincent Bonnet, Alejandro González. Determination of subject specific whole-body centre of mass using the 3d statically equivalent serial chain. 41(1):70–75. ISSN 09666362. doi: 10.1016/j.gaitpost.2014.08.017. URL https://linkinghub.elsevier.com/retrieve/pii/S096663621400681X.

[19] J. Wang and R. Lindenbergh. VALIDATING a WORKFLOW FOR TREE INVENTORY UPDATING WITH 3d POINT CLOUDS OBTAINED BY MOBILE LASER SCANNING. XLII-2:1163–1168. ISSN 2194-9034. doi: 10.5194/isprs-archives-XLII-2-1163-2018. URL https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLII-2/1163/2018/.