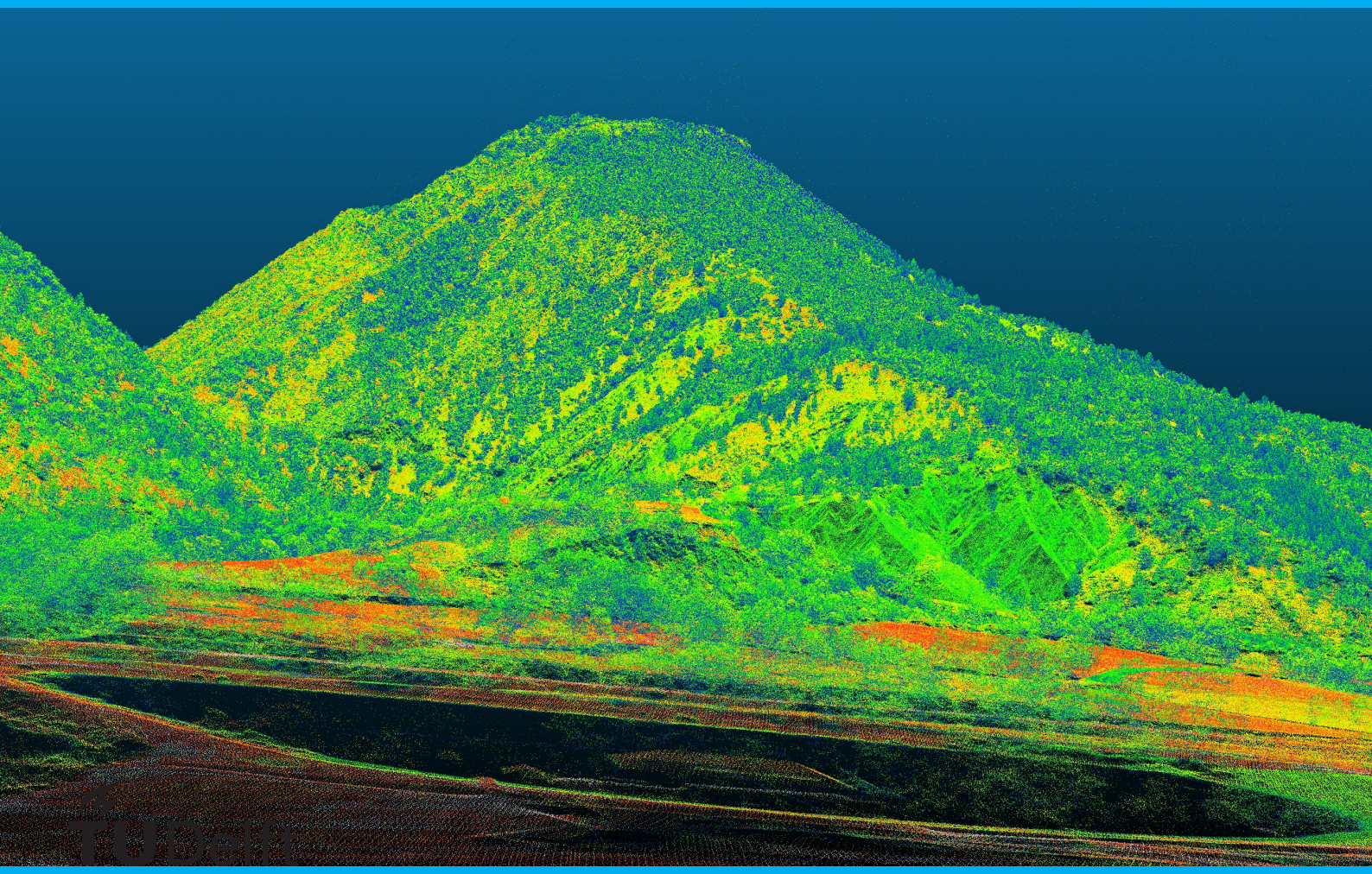
Airborne LiDAR data tree cover map

LiDAR data as additional data to Sentinel-2 spectral images to enhance land cover classification

Adam de Boer



Airborne LiDAR data tree cover map

LiDAR data as additional data to Sentinel-2 spectral images to enhance land cover classification

by

Adam de Boer

to obtain the degree of Bachelor of Science in Applied Earth Sciences
at the Delft University of Technology,
to be defended publicly on Wednesday January 18, 2023 at 01:00 PM.

Student number: 5150302
Project duration: November 14, 2022 – January 18, 2023
Thesis committee: Dr. R. C. Lindenbergh, TU Delft, supervisor
F. Dahle, TU Delft, supervisor

TU Delft

Preface

It was back in 2019, while sitting in a restaurant in Purmamarca, Jujuy, Argentina I decided to follow the Applied Earth Sciences program. After having visited an open day at the university in 2015, I thought it would be the best fit to keep travel opportunities open, have multicultural exchanges and be able to go outdoors.

And so, the journey began. The past 3.5 years have been pretty eventful; I have met nice people, quite quickly Covid-19 hit. The classes were taught online, and it gave a whole different dimension to studying. The pandemic offered nice times to think, to experience new kinds of working from home and it gave me the opportunity to regain the fun in cycling. It took long, I decided to spend some months in Sweden, where I studied online, learned Swedish and did sports.

Then classes started to be in person like a day a week, we started preparing for the fieldwork. I got to say, I considered it as the highlight of the program. Spending 2 weeks in La Drôme provençale, in Dutch I would say het is een Drôme om in la Drôme te zijn. The style of living, the Frenchies, the French language, it's beautiful down there in la douce France.

Fortunately, I was able to spend some more time in the Southeast of France. I did an exchange semester in Grenoble (38000), I followed classes with the program Sciences de la Terre (STE) and had some interesting courses like Glaciologie, where I would go on an excursion to la Mer de Glace in Chamonix. For Hydrologie, I would enter a small tributary river to l'Isère to measure the discharge while wearing fisherman clothes. Besides the courses, la Capitale des Alpes was paradise to me. You were literally next to la Chartreuse, le Vercors and la Belledone. You would just hike up a mountain within those massifs from the city, there were amazing camping opportunities and in winter you could go tour skiing. You would just bring your skis to school and along with the Frenchies we did go skiing after class. The people, from STE and Le Foyer Étudiant helped me improving French, taught me the game Présidents and we would interchange recipes. The French kitchen, la simplicité, the creaminess because of the amount of butter they use, it's delicious. Besides the people from STE and le Foyer, it was very nice meeting people coming from all over the place, especially la Banda de Alberto (Querer Es Poder) and los Obobos de Brati. If you are ever looking for a bar to spend some time, go to Brasserie la Natation, an exquisite atmosphere and it is in the same league as el Mono Blanco, in Arequipa, Peru. If you are ever in the Mono Blanco on a Monday night, say hi to Modesto, Leo, Nicole, Stephany, Denis, Patrick, Kevin, Jonathan, Steve y los demas de Intiwawa.

After the semester, I returned to Delft to follow some courses, I was training to run a marathon in Connemara with my mate Eric Sandía, I was too enthusiastic with training after my 30 k run, nevertheless I started the marathon while being injured. I couldn't walk afterwards, I had to learn to walk again. It was quite an experience. So was Grenoble, I didn't pass every course so I did an exchange semester at Utrecht University, I live in Utrecht now and have written the thesis. I chose the topic as I really liked the fieldwork in France and liked the task of making the land cover classification map. Sometimes it was tough to focus and from time to time I had a hard time working on the thesis. Nevertheless, the main goal was just to continue no matter the setback. I destroyed my keyboard during the process, did it help anything? No, I repaired it and continued. The small steps and continuous work delivered the result. Look, I can cry about the thesis being boring or not interesting, destroy keyboards, but it won't change the fact that I just got to do it. So, I just did it.

I am very thankful to Roderik and Felix, who were my supervisors for the thesis. Both Roderik and Felix were very enthusiastic in the project, they wanted to learn about it and it was a nice project to learn from the trees. Roderik and Felix helped me sending me in the right direction for the thesis, they would give inspiration and help send me in the right direction every week in and out. In addition to Roderik and Felix, I want to thank all the people affiliated with uni whom I experienced the journey with.

Besides, I want to thank the organisers of the Ogólnopolski Festiwal Kapel i Śpiewaków Ludowych in Kazimierz Dolny of 1992, where my parents met each other. Without that festival, I probably would not exist. I want to thank my parents, who have given me unconditional support all my life, during the studies and especially during the thesis to help motivate me finish the studies and always creating the opportunity to find my passions. I would like to thank my brother Stefan, who is the real programming king, whenever I got stuck programming, Stefan would find the mistake within the code in 5 minutes. We have been brothers for over 22 years, and even though we have very different personalities, I always admired you. I can't wait to get skiing again like we always have done.

The past half year in Utrecht has been a good half year to reflect on the studies, to look at future opportunities and write the thesis. What I have learned, is that I do not like spending lots of time programming on the computer and even though I study Earth Sciences, it's definitely not my passion. So what's my passion? I do not know, I know that I like to be in touch with people, get to know cultures, learn languages, I like to do sports and be outdoors. Combining those would vivir la vida mean for me. Hence, this journey at Delft is coming to an end. Even though I was not passionate for the studies, I had a fun time and I learned a lot from it. As Edith Pfiaf would say: "Non, je ne regrette rien". Now it's time to say goodbye, it's always tough to say goodbye, although it would be the next step in the development and as Gustavo Cerati would say: "Poder decir adiós, es crecer".

So, we march on like my grandfather would say: "siano, słoma, siano, słoma, siano, słoma", that's how I will start the next adventures. I can't, I really can't wait to go bikepacking after the thesis: De Boer goes on Tour!

Dziękuję za wszystko!

Adam de Boer
Delft, January 2023

Abstract

The following report will enlighten to which extent LiDAR data could enhance land cover classification. It focuses on the area Lemps (26510) in Southwest France where a land cover classification was made using Sentinel-2 spectral images during the fieldwork. Using the additional LiDAR data, the focus shifts to distinguish coniferous and deciduous trees. Training data from the LiDAR data has been selected in CloudCompare. Using the training data, features for coniferous and deciduous trees were extracted in python. The unique features were used as classifiers. Features based on the Intensity were found to be important. Based on the classifiers, two methods were used to classify the area. Random Forest and Nearest Neighbour were the classification methods. The classification using Random Forest was found to be more accurate. The Random Forest classification map has been compared with previously acquired Sentinel-2 classification maps. The Corine Land Cover Classification and the classification map from the fieldwork were compared to the classification of coniferous and deciduous trees using LiDAR. Lots of overlap was found with the Corine Land Cover, some overlap was present with the map acquired during the fieldwork. The map created during the fieldwork contained less training data, hence the model was not trained enough. If more training data is collected for both LiDAR and Sentinel-2 classifications, LiDAR data could enhance the general land cover classification. Especially taking the intensity into account as a classifier.

Keywords: LiDAR, classification, Random Forest, Nearest Neighbour, intensity

Contents

List of Figures

List of Tables

1 Introduction

1.1. Project Description

At the bachelor program of Applied Earth Sciences at the Delft University of Technology, students go on a fieldwork to Southwestern France. The goal is to get an in-field experience while creating a geological map and creating a land cover classification map. The land cover classification map has been determined by collecting training data and based on the training data, features were extracted from spectral satellite (Sentinel-2) data. Based on those features, several classification methods were used in the software QGIS to create land cover classification maps. Recently, airborne LiDAR data has been made public on the region where the fieldwork was conducted. LiDAR data is a point cloud dataset obtained with a laser system. LiDAR data could be used as a base layer to extract features for the training data and hence could be used to improve the classification.

1.2. Problem Statement and Research Question

The aim of this report is to qualify the use of LiDAR data as base data for land cover classification. The land cover classes, that will be looked at, are coniferous and deciduous trees. Characteristics of coniferous and deciduous trees will be derived using python. The unique characteristics are features. With help of those features, a classification will be made.

The main research question is:

To which extent can LiDAR data be used to enhance the land cover classification for coniferous and deciduous trees?

Sub-questions supporting the main research question are:

- *How can the LiDAR data be acquired and processed?*
- *How can features be derived from the LiDAR data and which features can be applied?*
- *To what extent are the extracted features significant?*
- *How can the significant features be used for the classification?*
- *To what extent compare the LiDAR classified trees to the Sentinel-2 classified trees?*

1.3. Thesis Structure

The report will be presented in the following structure. Chapter 2 will give background information on the area, trees, Sentinel-2, classification and an introduction to LiDAR will be given. Subsequently, Chapter 3 shows a step by step methodology on how to create the classification map using LiDAR. In Chapter 4 the classification maps of the trees are shown and compared to Sentinel-2 classification maps. The report will be finalised with the conclusion and recommendations in Chapter 5. The

processing of the LiDAR data and its modelling has been conducted in softwares such as QGIS, CloudCompare, Google Earth, Google Maps and the programming language python is used. An overview of the thesis structure is shown in Table 1.1.

Table 1.1: Structure of Report

Chapter #	Title of Chapter
2	Background Information
3	Methodology
4	Results
5	Conclusion and Recommendations

2

Research Area Characteristics

Within this chapter, the area of research will be shown. It will be followed by an overview of the trees in the area. An explanation on Sentinel-2 will be given before the term classification will be explained. Subsequently, the previous acquired classification map using Sentinel-2 data will be presented (Figure 2.10). It will be concluded by information on LiDAR.

2.1. Area

The area studied for the project during the fieldwork is in the French department La Drôme. The studied area is called Lemps (Blom, 2021) and ranges between the coordinates [886000,894000,6362000,6368000] in the coordinate system (EPSG:2154). Thus the fieldwork area comprises of 48 square kilometers. Within the area, this project will focus on six areas of one km^2. The area is shown in Figure 2.1.

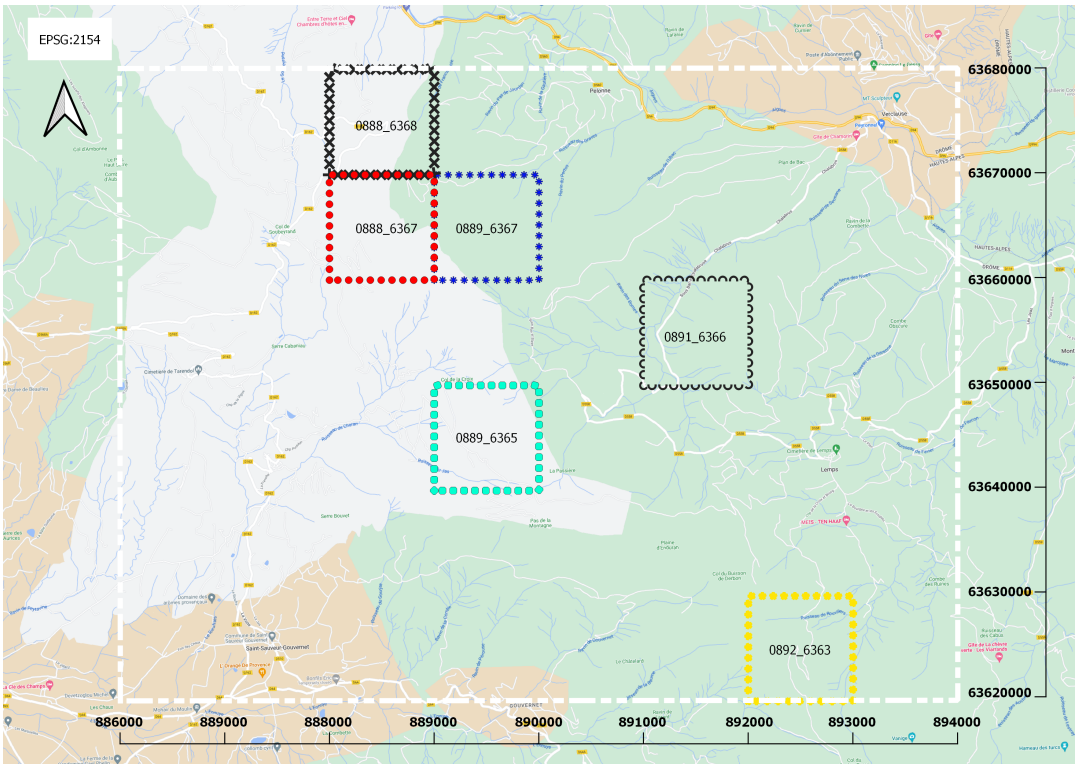


Figure 2.1: The fieldwork area of *Lemps* with the areas that will be classified

2.2. Trees

Within this section, the trees in the area will be described and the distinction between coniferous and deciduous trees will be made.

2.2.1. Trees in the region

The most common trees to occur in the region are the Pinus sylvestris, Quercus pubescens, Fagus sylvatica and the Abies alba. Those trees account for 68.2% of the trees in the region between 2005 and 2014 (Auvergne-Rhône-Alpes, 2019). This means two types of coniferous trees are present: Pinus sylvestris (2.2.2) and the Abies alba (2.2.2) as well as two types of deciduous trees: the Quercus pubescens (2.2.3) and the Fagus sylvatica (2.2.3).

2.2.2. Coniferous Trees

Within the subsection, the Pinus Sylvestris and the Abies alba are described. In Figure 2.2 the two trees are visualised.

Figure 2.2: The Pinus sylvestris on the left side (“Pinus sylvestris”, 2022) and the Abies alba on the right side (“Abies alba”, 2022)

Pinus sylvestris

The Pinus sylvestris, Scots Pine, forms part of the Pinaceae family. The tree is native to Eurasia, its main characteristics are that the leaves are blue-green and fairly short. Besides, its bark is orange-reddish. It can be 35 metres high and usually has a lifespan between 150-300 years (Farjon, 2017).

Abies alba

The Abies alba, the European silver fir or silver fir, is a tree from the Pinaceae family as well. It is a native tree species to Europe, spread from the Pyrenees to the Balkan and Romania. The trees can grow up 40 to 50 metres high and the trunk can be 1.5 metre thick. Usually, the tree grows on an altitude between 300 - 1700 metres above sea level. To survive, the tree requires at least 1,000 millimetres of rainfall annually. The leaves are needle-like, up to 3 cm long and 2 mm wide and the cones are usually 9-17 cm long (Farjon, 2017).

2.2.3. Deciduous Trees

Within the subsection, the Quercus pubescens and the Fagus sylvatica are described. In Figure 2.3 the two trees are visualised.

Figure 2.3: The Quercus pubescens on the left side (“Quercus pubescens”, 2022) and the Fagus sylvatica on the right side (“Fagus sylvatica”, 2022)

Quercus pubescens

The Quercus pubescens, the downy or pubescent oak, is a species of white oak. It is spread between the Pyrenees, Crimea and the Caucasus. The tree is a medium-sized deciduous tree, and can grow up to 20 metres tall. Its twigs are light purple or whitish. The leaves are between 4 and 10 centimetres long and 3 to 6 centimetres wide. Beyond the middle of the tree the leaves are the widest (Rushforth, 1999).

Fagus sylvatica

The Fagus sylvatica, the European or common beech, is a deciduous tree. It can reach heights up to 50 metres and its trunk can have a 3 metre diameter. The leaves are between 5 to 10 centimetres long, and can be 3 to 7 cm wide. It is a tree which has it origins from the European mainland (Durrant et al., 2016).

2.2.4. Location of trees in the region

The distribution of the trees in the area is visualised by the Corine land cover classification. The land cover map has a resolution of 100 metres and 48 classes are defined for 39 countries in Europe. Within the the area of interest, the land cover classification contains 13 classes. From those 13 classes, three classes are related to trees and those are displayed in Figure 2.4. The tree classes are *311-Broad-leaved forest, 312-Coniferous forest and 313-Mixed Forest*. From which 311-Broad-leaved-forest are the deciduous trees, 312-Coniferous forest are coniferous trees and 313-Mixed Forest are coniferous and deciduous trees (“Corine Land Cover”, 2018).

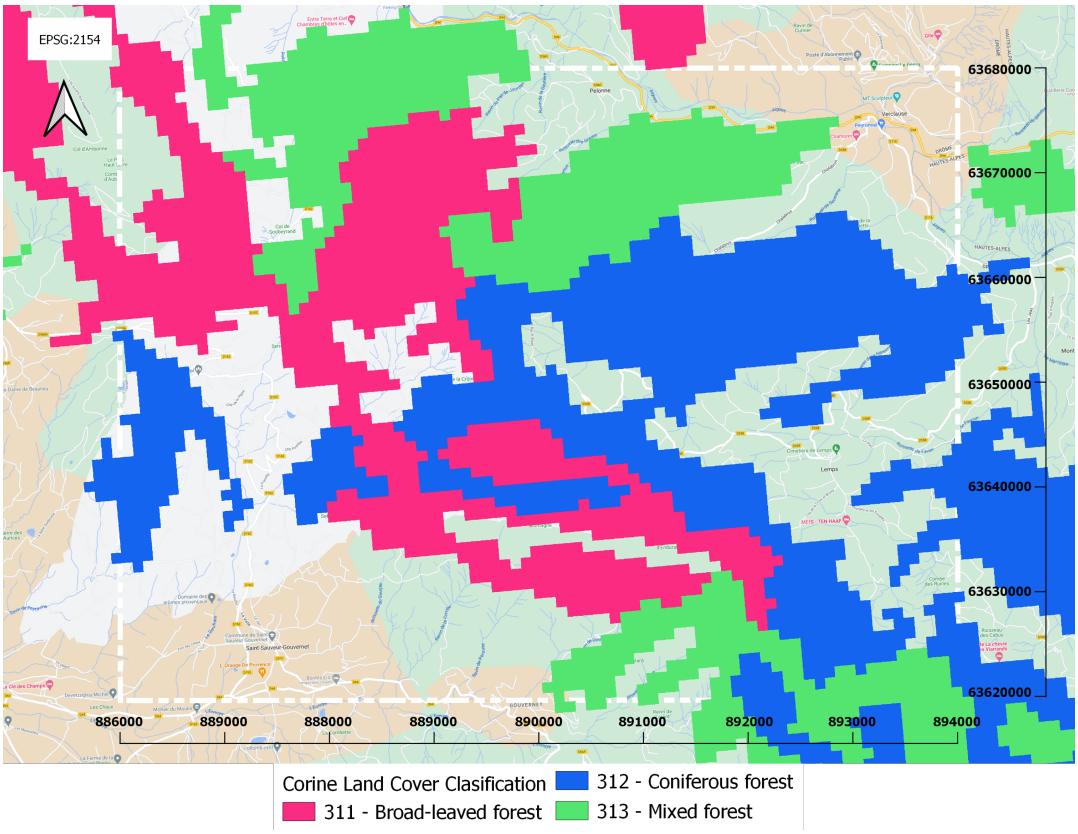


Figure 2.4: The area of *Lemps* shown overlaid with trees in the Corine land cover classification

2.3. Sentinel-2

The goal of Sentinel-2 mission launched by the European Space Agency is to acquire high-resolution and multi-spectral images. Those are being used for multispectral observations and applications such as land management, agriculture and forestry (ESA, n.d.-b). The satellites have a temporal resolution of five days at the equator if two satellites are used and 10 days if one satellite is used. Currently, two satellites are in use since 2017 (ESA, n.d.-b). The satellites contain a multispectral instrument which measures the reflected radiance in 13 bands. Four bands have a spatial resolution of 10 metres, six bands at 20 metres and three bands at a spatial resolution of 60 metres. The spectrum from the bands goes from Very Near Infrared to Shortwave Infrared wavelengths. The bandwidth is measured at Full Width Half Maximum leading to the calculation the central wavelength at the barycentre of the spectral response function leading to the Equation 2.1:

$$\lambda_c = \frac{\int \lambda \times S(\lambda) \mathrm{d}\lambda}{\int S(\lambda) \mathrm{d}\lambda} \tag{2.1}$$

Where λ is the wavelength, $S(\lambda)$ is the instrument spectral response function and λ_c is the central wavelength of a given spectral band. The wavelengths are measured in nm.
Those bands give a different response to every wavelength as can be seen in Figure 2.5. A different scatter is present for every colour band, an image is generated using 3 colour bands at maximum. Because every band has a different response scatter, multiple combinations can be made for different applications such as vegetation, agricultural cover, geology, etc. (ESA, n.d.-a).

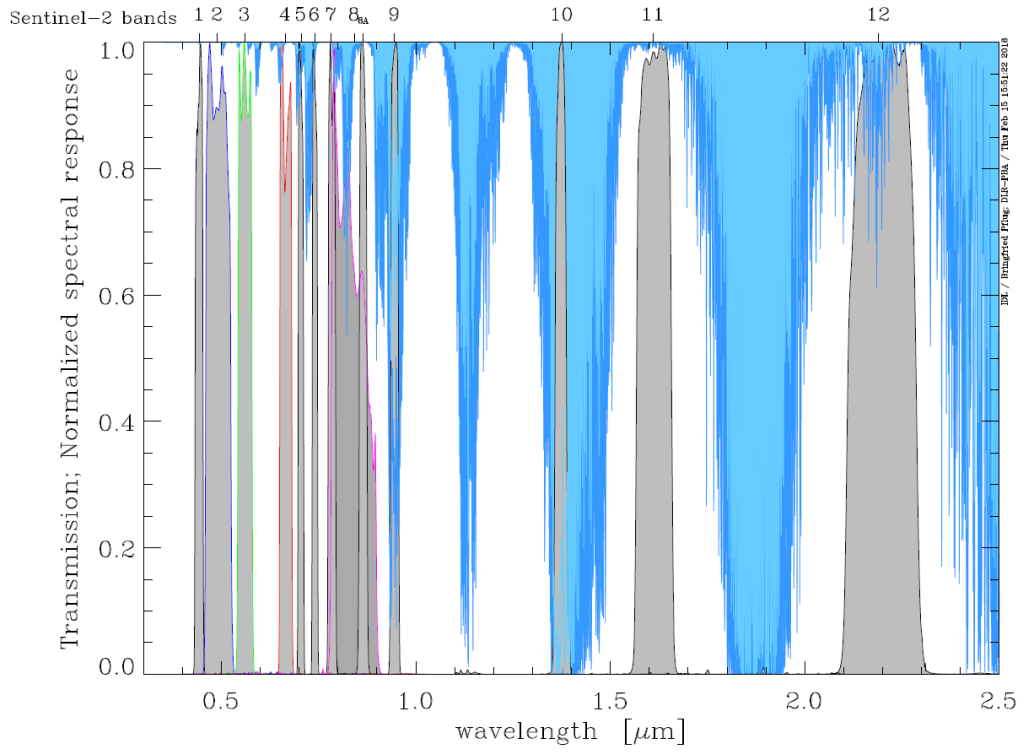


Figure 2.5: Spectral Response Functions from S2A and transmission due to Water Vapour absorption (ESA, n.d.-a).

2.4. Airborne LiDAR data

This section will give an overview of Airborne LiDAR data, will briefly introduct the data used for the report and point clouds will be shown for coniferous (Figure 2.7) and deciduous trees (Figure 2.8).

2.4.1. LiDAR in general

LiDAR is a remote sensing surveying method and its name is derived from Light Detection And Ranging. It measures the distance to a certain point. A light pulse is sent out to a certain point. As soon as the emitted light gets reflected by an object and will be detected by the reflector adjacent to the transmitter. The travel time from transmitter to reflector and back will be divided by 2. This travel time, along with the speed of light, will be used to calculate the distance between transmitter and reflector as shown in Equation 2.2 (Lindenbergh, 2021). The visualisation of the travel time is calculated is displayed in Figure 2.6.

$$d = \frac{1}{2} \cdot c \cdot t \tag{2.2}$$

Where d is denoted as the distance between transmitter and reflector, c is the speed of light equal to $3.0 \cdot 10^8\ m/s$ and t is the two-way travel time divided over 2, as can be seen in the Figure 2.6.

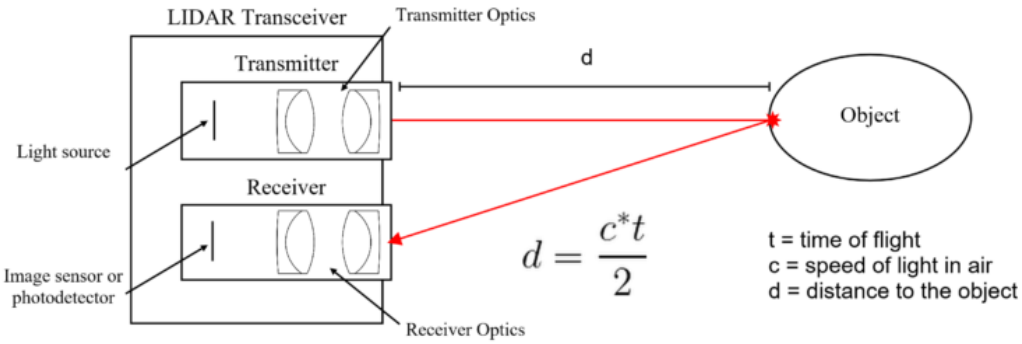


Figure 2.6: Visualisation of how LiDAR works (Optics, n.d.).

2.4.2. LiDAR data in France

The LiDAR data in France is collected by attaching a laser system is attached to an air vehicle like an aeroplane or helicopter. The system sends out a laser light and captures its reflection to calculate the distance to the certain point. Along the laser system, a GPS and an Inertial Measurement Unit (IMU), which records acceleration and rotation data, are installed to obtain the flight trajectory of the aeroplane. Thus the measured distances could be assigned to a location (Vosselman and Maas, 2010).

The LiDAR data that became available in the fieldwork region became available in November 2022 through the Institut Géographique National. The LiDAR data is measured in 2021 and is divided into tiles of one by one kilometer. Hence, for the area 48 tiles could be used. Within the point data, 16 variables are measured and are shown in Table 2.1 and will be explained below.

Table 2.1: Variables measured with LiDAR

X	Y	Z	Intensity
Return Number	Number of Returns	Scan Direction Flag	Edge of Flight Line
Classification	Synthetic	Key Point	Withheld
Scan Angle Rank	User Data	Point Source ID	GPS Time

Variables from upper row in Table 2.1

The measured *X* and *Y* return the X- and Y-coordinate in EPSG:2154. *Z* is the altitude with respect to the sea level. The *Intensity* shows the strength of the reflected light pulse sent out by the laser system.

Variables from second row in Table 2.1
The *Return Number* is the pulse return number for a given output pulse, the laser pulse can have up to 5 returns, and the returns are numbered in their sequence of return, the dimension *Number of Returns* is the total number of returns for a given pulse, for instance it might be return of 2 within a total of 5 returns (ASPRS, 2008). The *Scan Direction Flag* is a value for the direction of the scanner mirror, which could be either 0 or 1. If the scanner mirror is located to the left with respect to the aeroplane, the value is negative and corresponds to value 0. It is 1 as the mirror is located to the right, and thus positive. The *Edge of the Flight Line* dimension would return a value of 1, at the last measured point before the aeroplane changes its direction.

Variables from third row in Table 2.1
Classification returns the classification of the area, in the case of this data, it always returns a value of 1, which means unclassified. *Synthetic*, means that a point in the cloud was acquired in a different way than LiDAR; *Key Point*, denotes a point in the cloud which is key for classification and *Withheld* means a point should be deleted.

Variables from last row in Table 2.1
The *Scan Angle Rank* denotes the direction of scanning when the laser was used, it ranges from $-90°$ to $90°$, measuring from left to the right with respect to the laser carrier. *User Data* is data defined by the user, *Point Source ID* is the ID of a point and *GPS Time* measures the time from the measuring device at which the point was measured (ASPRS, 2008).

Trees visualised in point clouds
The LiDAR data for the trees is visualised for coniferous trees in Figure 2.7 and for deciduous trees in Figure 2.8.

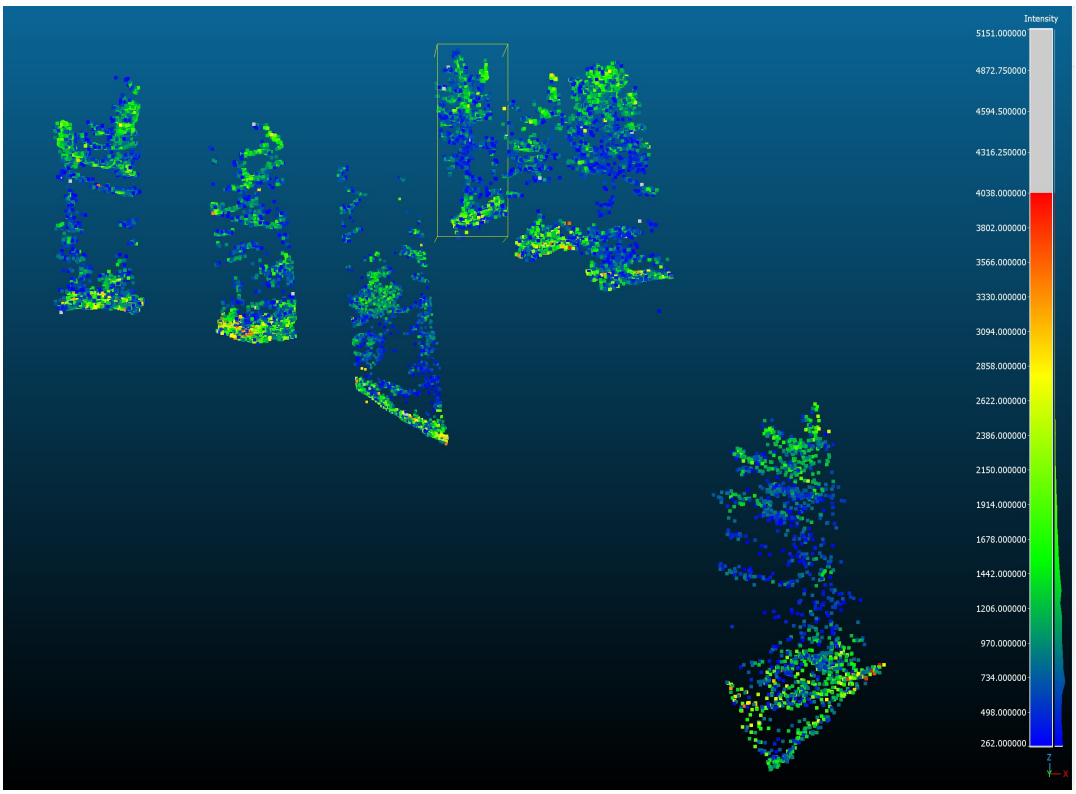


Figure 2.7: Coniferous trees Point Cloud Data visualised in CloudCompare

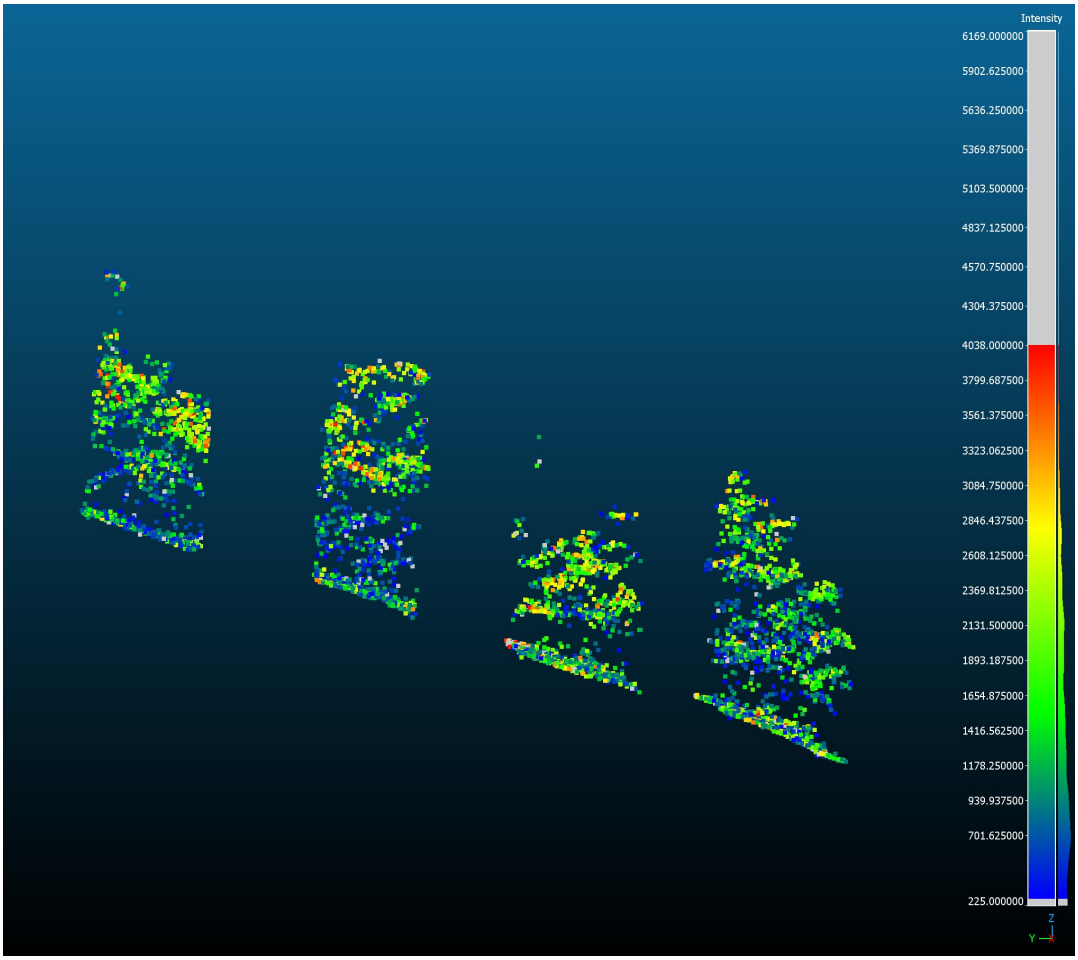


Figure 2.8: Deciduous trees Point Cloud Data visualised in CloudCompare

2.5. Classification Methods

This section will define classification, presents the classification methods used and it will show the fieldwork classification map in Figure 2.10.

2.5.1. Classification

Classification is the process of ordering objects in a group based on diagnostic criteria (Sokal, 1974). Training data needs to be selected from a base layer on which the classification will take place. Some characteristics of the coniferous and deciduous trees need to be extracted from the training data, this is done by processing. Those unique features, will be the classifiers for the model. The area, that will be represented, will be characterised using the classifiers. Thus, a classification map is made.
Different classification methods might be used, during the fieldwork a classification map was made using an image from Sentinel-2 as a base layer. Three methods were applied: Nearest Neighbour, Maximum Likelihood and Spectral Angle Mapping. For this project, a base layer is used from Airborne LiDAR data and two classification methods were used: Nearest Neighbour and Random Forest.

2.5.2. Nearest Neighbour

The Nearest Neighbour classification method is a method which does not require lots of computational power. It is a simple classification method where the value of the site, that will be classified, is determined by duplicating the value of its nearest site. The value is based on the properties of the data, so if the area shows similar properties as the classified area, it will duplicate that value. If the to be measured site is close to multiple observations, the average of those values is calculated. As the Nearest Neighbour classification method duplicates the nearest value, the result might be blocky (Li

et al., 2021). If the classification is based on non-numerical values, a value can be assigned to the non-numerical values. If there is a relationship between the non-numerical values, the same value can be assigned to the non-numerical value (Luo et al., 2020).

2.5.3. Maximum Likelihood

The Maximum Likelihood classification method decides the class based on which class has the biggest likelihood for a certain location. The algorithm is built on the concepts of the probability distribution based on the Bayes' theorem. A Multivariate Gaussian distribution is used, and the class with the highest probability will be selected. The function is calculated for every pixel by Equation 2.3 (Congedo, 2021).

$$g_k(x) = \ln p(C_k) - \frac{1}{2}\ln\left|\sum_k\right| - \frac{1}{2}(x - y_k)^T \sum_k^{-1}(x - y_k) \tag{2.3}$$

Within Equation 2.3, the C_k denotes the land cover class, x the spectral signature vector of an image pixel, pC_k the probability that the assigned land cover class is the actual land cover class. $\left|\sum_k\right|$ is the determinant of the covariance matrix of the data in class C_k. $\sum_k^{-1}$ is the inverse of the covariance matrix and y_k is the spectral signature class of class k (Congedo, 2021).

2.5.4. Spectral Angle Mapping

The Spectral Angle Mapping classification algorithm was used for the fieldwork classification. The method calculates the spectral angle between spectral signatures of the training data and the spectral signature of the image pixels. The spectral angle θ is defined in Equation 2.4 (Congedo, 2021).

$$\theta(x, y) = \cos^{-1}\frac{\sum_{i=1}^{n} x_i y_i}{(\sum_{i=1}^{n} x_i^2)^{\frac{1}{2}} * (\sum_{i-1}^{n} y_i^2)^{\frac{1}{2}}} \tag{2.4}$$

In Equation 2.4, x denotes the spectral signature vector of an image pixel, y denotes the spectral signature vector of the training area, n is the number of spectral bands, in the case of Sentinel-2 data, n is equal to 13. Thus, the pixel belongs to the class having the lowest angle, which equals to Equation 2.5.

$$x \in C_k \iff \theta(x, y_k) < \theta(x, y_j) \forall k \neq j \tag{2.5}$$

Where C_k denotes the land cover class, y_k the spectral signature of class k and the spectral signature class of j is denoted by y_j.

2.5.5. Random Forest

The Random Forest classification method is based on multiple decision trees. In every decision tree a choice is made for the output based on the input data. Based on the majority vote, a class is assigned. Within the decision trees, a gini impurity is present. It ranges between zero and one, if it is zero it indicates perfect classification and it selects always the right class. If the gini impurity is one, it always predicts the false class (Dimitriadis et al., 2018). An example of a decision tree, which is not used in the final classification, is presented in Figure 2.9.

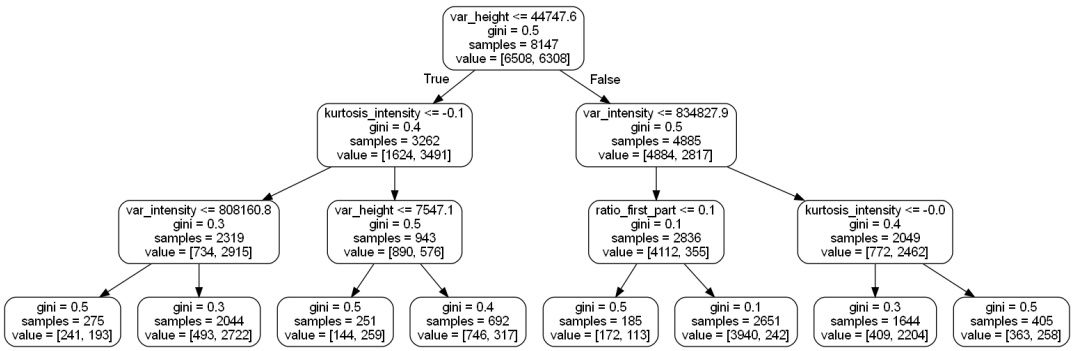


Figure 2.9: Example of a decision tree for classifying using the Random Forest classification method

2.5.6. Fieldwork 2021 - Spectral Angle Classification

During the 2021 Fieldwork, a land cover classification map has been made in QGIS. Using the Sentinel-2 derived data, along with the training data, the area has been classified using the Semi-Automatic Classification plugin. This is a plugin in QGIS that allows for supervised classification of remote sensing images (Congedo, 2021). For the classification, several methods have been used such as Nearest Neighbour, Maximum Likelihood and Spectral Angle Mapping. The spectral angle classification method was selected as it was the most stable during the classification and is shown in Figure 2.10.

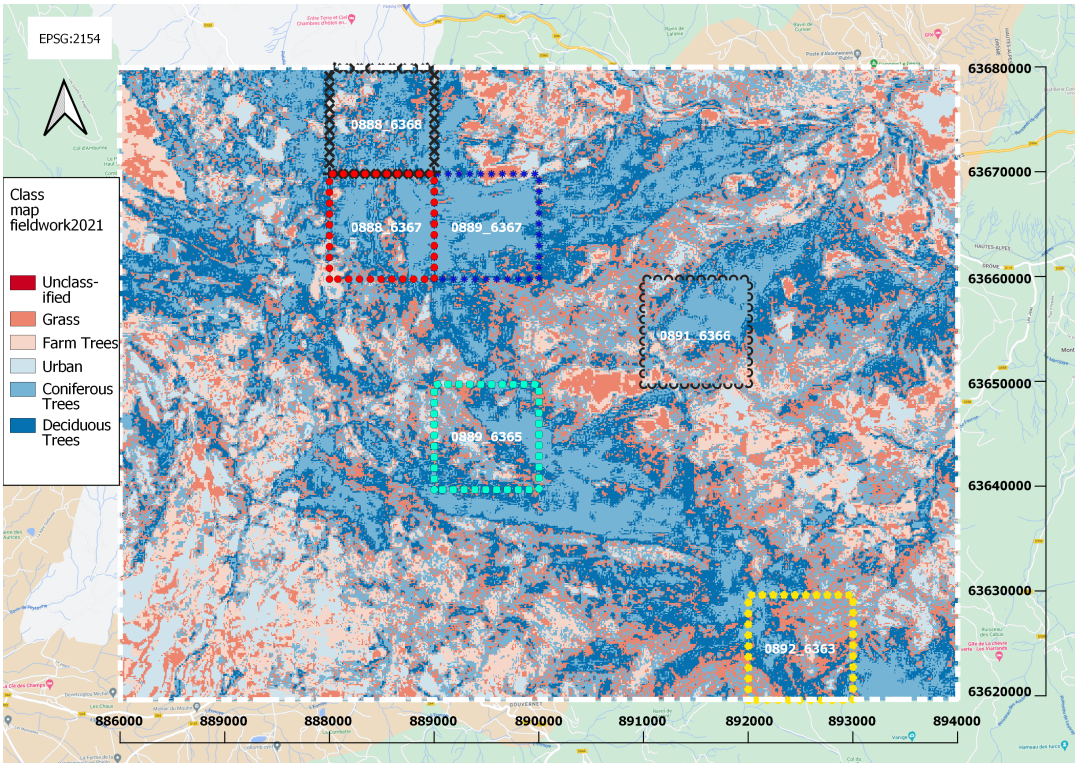


Figure 2.10: Classification map obtained during fieldwork 2021

Within the classification map obtained in 2021, six classes were defined: Unclassified, Grass, Farm Trees, Urban, Coniferous Trees and Deciduous Trees. Six were chosen as the fieldwork had to be conducted in two weeks, and more classes would need more time for analysis. For this project, the focus is on to distinguish the coniferous and deciduous trees. Therefore, the area with just the trees are shown in Figure 2.11.

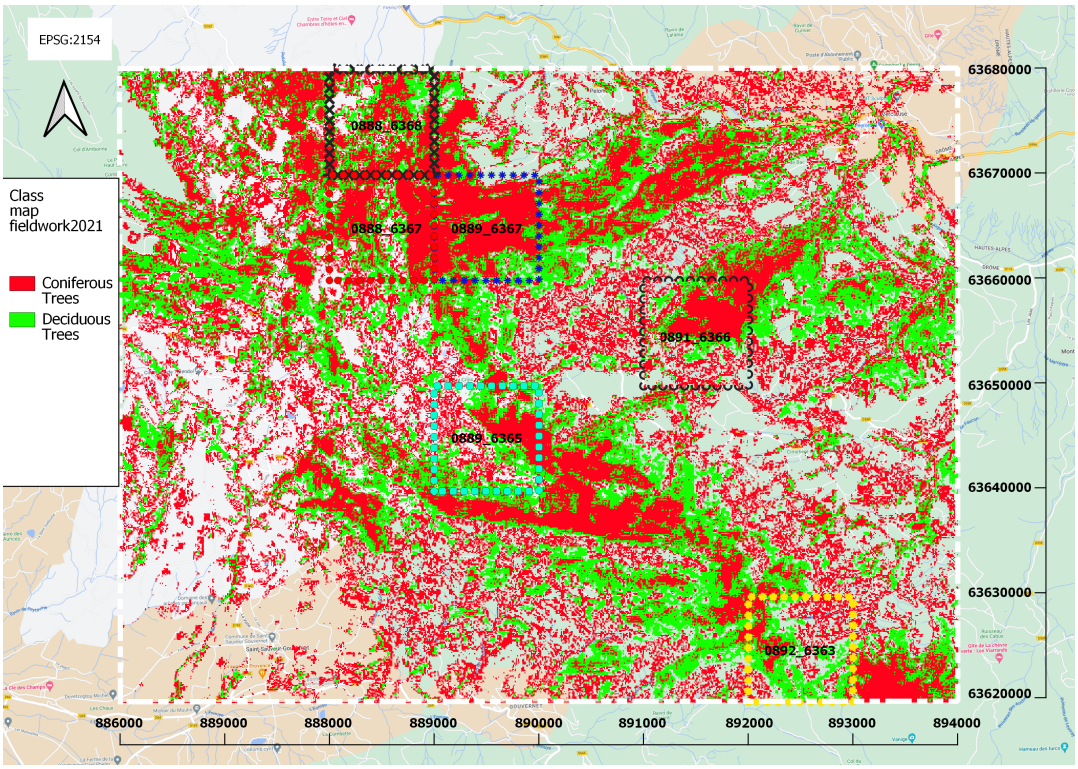


Figure 2.11: Classification map obtained during fieldwork 2021, with only Coniferous and Deciduous being displayed

3

Methodology

In the chapter Methodology, an overview will be given on the methods displayed in Figure 3.1.

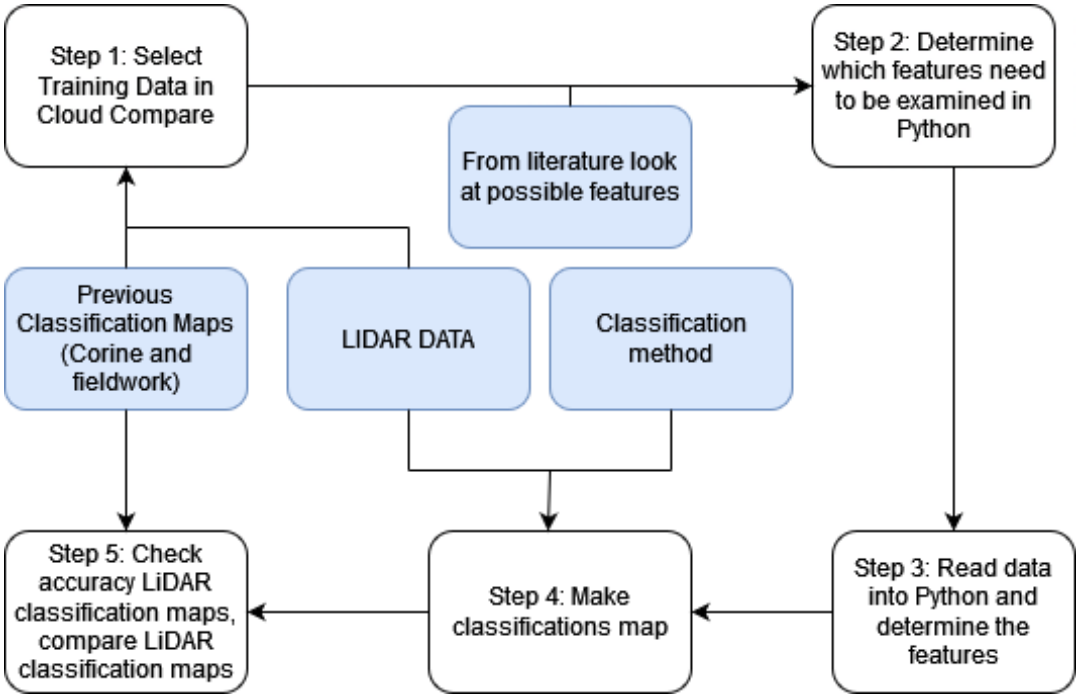


Figure 3.1: Workflow of the report

3.1. Step 1: Select Training Data in Cloud Compare

Within the first step of the project. Training data needs to be selected from the area. Using CloudCompare, 10 blocks for both coniferous and deciduous trees have been selected based on previous classification of coniferous and deciduous trees in the Corine land cover classification map (Figure 2.4) and the fieldwork 2021 classification map (Figure 2.10). Those blocks have been divided afterwards in smaller patches using Python. The average area of the patches was $8.01m^2$ and $8.61m^2$ for coniferous and deciduous trees respectively. In total, 17093 patches of data are formed, from which 12816 are used as training data and 4273 patches as validation data. The tiles used for the training data are *0886_6366*, which contains both deciduous trees and some coniferous trees. The tile *0888_6367* was used for the retrieval of deciduous trees training data. More coniferous trees training data was selected from the tile *0891_6366*. The selected training data is shown below in Figure 3.2.

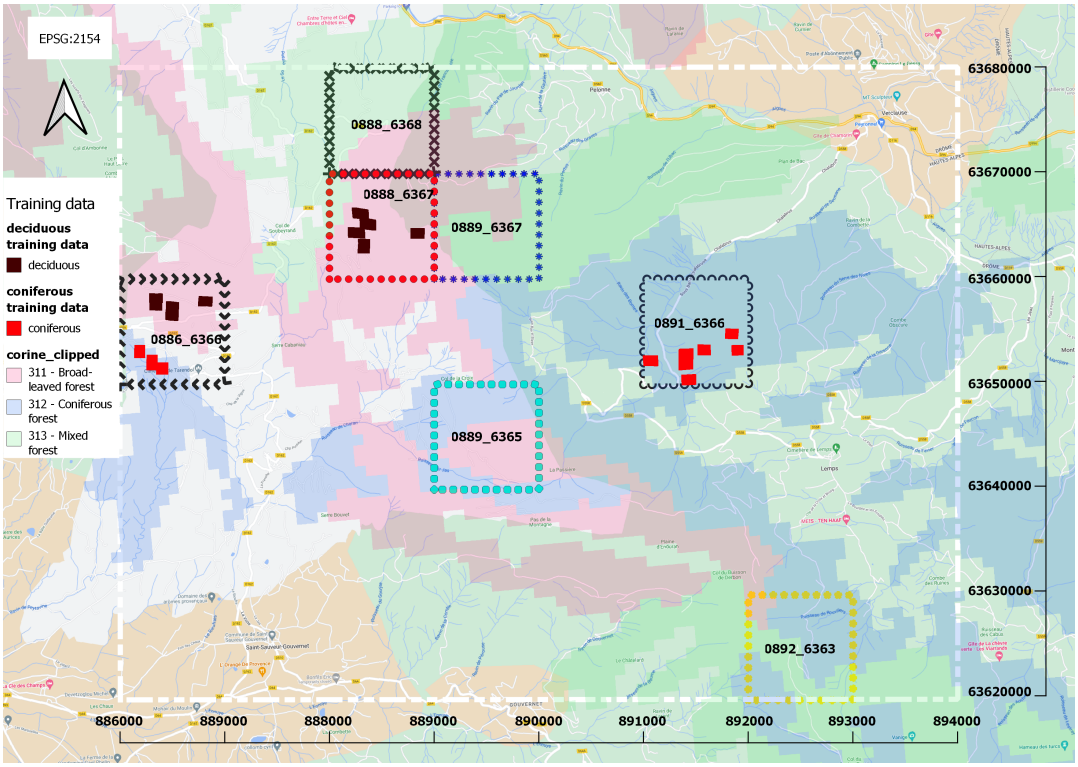


Figure 3.2: Training data with underlying Corine land cover classification (Figure 2.4) as a baselayer

3.2. Step 2: Determine which features need to be examined in Python

Coniferous and deciduous trees have features. Those who differ could be used as a classifier in classification. Within this step, those features, that differ, will be explained (Table 3.1). The features are based on literature from Labaar, 2022 and Heywood, 2020.

Table 3.1: Features that were examined at initial feature extraction

Features
Relative Height Average
Relative Height Standard Deviation
Kurtosis Height
Skewness Height
Variance Height
Coefficient of Variation Height
Average Intensity
Standard Deviation Intensity
Kurtosis Intensity
Skewness Intensity
Variance Intensity
Coefficient of Variation Intensity
Ratio 0 - 25% Height
Ratio 25 - 50% Height
Ratio 50 - 75% Height
Ratio 75 - 100% Height
Lower Half Ratio (0 - 50%) Height
Upper Half Ratio (50 - 100%) Height
Average Number of Returns
Average Return Number

Relative Height and Intensity Average

Firstly, the average height of a tree group is calculated. The elevation of the datapoints is denoted by the *Z-dimension*. Usually the average would be taken of the whole dataset. In this case, the relative height is used as the area is mountainous. Therefore taking the height at a certain location could lead to bigger height difference than the actual tree height. The relative height is calculated by subtracting the the lower five percent from the upper five percent (Equation 3.1). Using the relative height differences, the average was calculated by adding all the differences, and then dividing those over the total number of differences, n, as can be seen in Equation 3.2.

$$Rel\_height = (Z \geq 95\%) - (Z \leq 5\%) \tag{3.1}$$

$$rel\_height\_avg = \frac{Rel\_height_1 + Rel\_height_2 + ... + Rel\_height_n}{n} \tag{3.2}$$

The Equation 3.2 could be used to calculate the average intensity as well, the relative height is changed by the intensity in the formula. The intensity is calculated for the whole dataset and not for a fraction like the relative height.

Relative Height and Intensity Standard Deviation

The standard deviation is calculated for both the relative height and the intensity. The standard deviation shows the variation about the mean and is denoted by Equation 3.3. Where n is the total number of data inputs.

$$std = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} \tag{3.3}$$

Kurtosis
In general, the kurtosis is a measure of the distribution's tail. The tailedness depicts how often outliers occur. Excess kurtosis is the tailedness of a distribution relative to a normal distribution. The kurtosis can be defined into three groups. It could be negative, zero or positive for which the distributions would be *platykurtic, mesokurtic and leptokurtic* respectively (Zbigniew, 2017). An example is shown below Figure 3.3. For every data patch the kurtosis is calculated and thus is either negative, zero or positive.

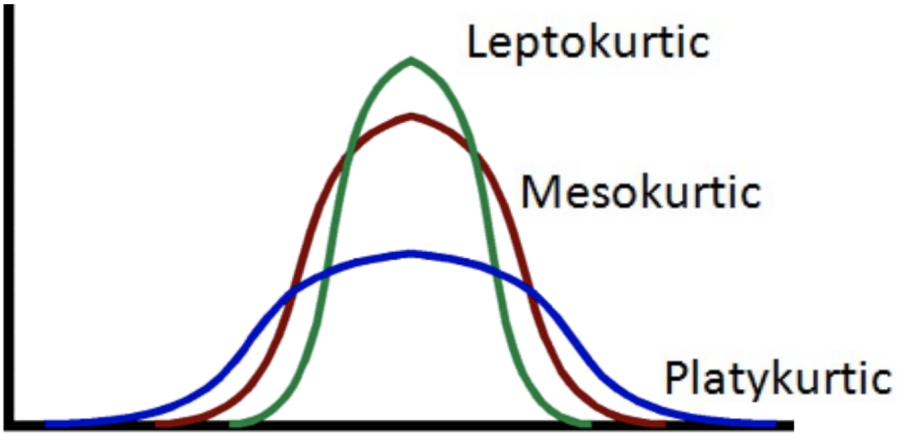


Figure 3.3: Platykurtic, mesokurtic and leptokurtic distributions (Donges, 2019)

Skewness
The skewness of a distribution describes the (a)symmetry of a distribution and is used for both the height and intensity. When the skewness is negative, the graph is skewed to the left and thus the left tail is bigger. If the graph is positively skewed, the tail of the distribution is bigger to the right side. For a visualisation of the skewness please refer to Figure 3.4. The skewness is computed as the Fisher-Pearson coefficient of skewness as can be seen in Equation 3.4 and Equation 3.5, where N is the total number of entries within the data patch. $\overline{x}$ is the average value and $x[n]$ is every entry within the data patch (Zwillinger and Kokoska, 2000).

$$skewness = \frac{m_3}{m_2^{\frac{3}{2}}} \tag{3.4}$$

$$m_i = \frac{1}{N}\sum_{n=1}^{N}(x[n] - \overline{x})^i. \tag{3.5}$$

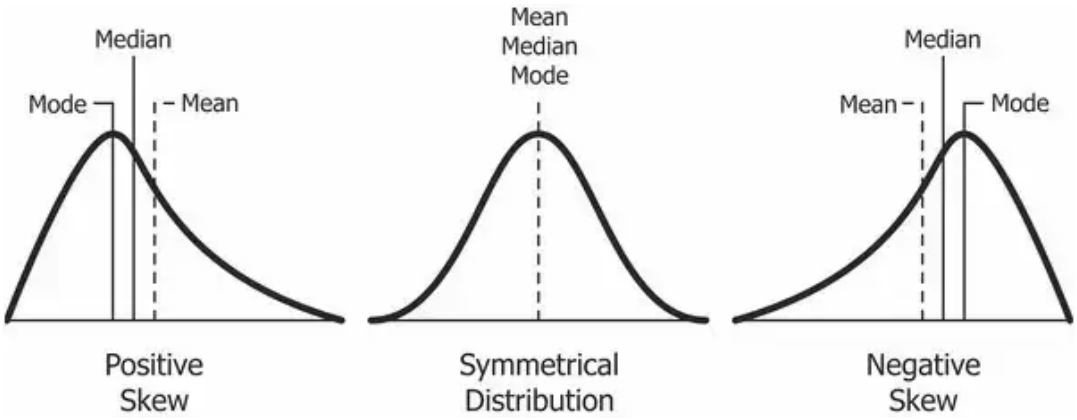


Figure 3.4: Positive skew, symmetrical distribution and negative skew (Donges, 2019)

Variance

The variance describes the spread of a distribution, if there is a high variance within the distribution, the data is more spread. The variance is described by the following equation Equation 3.6, the standard deviation is squared.

$$var = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} \tag{3.6}$$

Coefficient of Variation

The coefficient of variation is the ratio of the standard deviation to the mean, see Equation 3.7.

$$CV = \frac{std}{avg} \tag{3.7}$$

Height Ratios

For the height, the ratio of point distribution will be looked at. The point at the ground, is the lowest point and defined as the zero percent, the highest point has the 100% value. Thus, six ratios will be analysed, the ratio from 0 - 25%, 25 - 50%, 50 - 75%, 75 - 100%, 0 - 50% (lower half ratio) and 50 - 100% (upper half ratio). The ratio for 0 - 25% is defined in Equation 3.8, where the total number of height points in the interval gets divided over the total number of height points.

$$Ratio\ (0-25\%) = \frac{number\ of\ height\ points\ in\ ratio\ (0-25\%)}{total\ number\ of\ points} \tag{3.8}$$

Average Number of Returns and Return Number

As already has been mentioned in LiDAR data in France, the Return Number is how often the light pulse hits an object and get returned. the Number of Returns defines how often the actually returned number is a different part than the previous return(s).

3.3. Step 3: Read data into Python and determine the features

Subsequently, the training data has been loaded into Python. As already has been mentioned in subsection 2.4.2, 16 variables have been measured for the LiDAR data. Six variables are used for the calculations. The modified data has been stacked into $n \times 6$ arrays, with n denoting the number of points, which is different for every dataset. 6 denotes the number of columns. The six columns are *X, Y, Z, Intensity, Number of Returns and Return Number*. Afterwards, the data has been processed to calculate the values for the features described in section 3.2. The relevant features can be used for classification. The updated features are displayed in Table 3.2 and the results are presented in Table B.1 (Appendix B).

Table 3.2: Features that were based on training data

Features
Relative Height Average
Skewness Height
Average Intensity
Kurtosis Intensity
Ratio 0 - 25 % Height
Lower Half Ratio (0 - 50 %) Height
Upper Half Ratio (50 - 100 %) Height
Average Number of Returns
Average Return Number
Variance Height
Variance Intensity

3.4. Step 4: Make classifications map

At first, a classification map is made using the Random Forest Classification method. To create the map for the six areas, the RandomForestClassifier from scikit-learn library has been used. 100 decision trees have been used and no max depth of the tree set, which means that the calculations will be done until a gini impurity of zero is reached. Eleven features are used for the Random Forest Classification maps. Which means, in every decision trees $\sqrt{11}$ features are used. Using the RandomForestClassifier method, the importances of the features can be derived and are presented in Table 3.3.

Table 3.3: Importances of the features in percentages using RandomForestClassifier

name	rel height avg	skew-ness height	avg in-ten-sity	kurto-sis in-ten-sity	ratio first part (0-25%)	lower half range ratio	upper half range ratio	avg num ret	avg ret num	var height	var in-ten-sity
0888 6367	5.60	4.26	17.48	12.64	9.19	3.56	3.71	4.71	4.34	6.32	28.20
0888 6368	5.54	4.04	15.85	11.92	8.57	3.45	3.64	5.09	4.39	5.72	31.80
0889 6365	6.09	4.10	14.65	11.47	8.84	3.57	3.55	4.52	4.37	6.04	32.78
0889 6367	5.50	4.60	15.97	11.44	9.38	3.04	3.39	5.13	4.18	6.33	31.03
0891 6366	5.45	4.18	14.50	13.41	8.82	3.41	3.40	4.84	4.25	6.15	31.58
0892 6363	6.19	4.39	15.54	13.43	8.56	3.70	3.70	4.45	4.35	6.03	29.67

Within Table 3.3, a green color is given for every input which has an importance of more than five percent. From the results, it is clear that features based on the intensity are very important. Based on the importance, the features get updated for the Nearest Neighbour classification method, the updated features are represented in Table 3.4.

Table 3.4: Features that were based on RandomForestClassifier importances

Features
Relative Height Average
Average Intensity
Kurtosis Intensity
Ratio 0 - 25% Height
Variance Height
Variance Intensity

Classification maps have been made for Nearest Neighbour classification method using the 6 features (Table 3.4). Every data patch will be assigned to a class having its features most similar to the training data. The kNearestNeighbour from sci-kit learn has been used for the implementation, taking into account five neighbours, thus duplicating the value based on the majority vote of the 5 neighbours.

3.5. Step 5: Check accuracy LiDAR classification maps, compare LiDAR classification maps

For the final step, the accuracy for the method is calculated using the confusion matrix and the classified maps will be compared. A confusing matrix compares, class by class, the reference data and the validation data. Using a confusing matrix, the producer's accuracy and the user accuracy can be measured for each class. Thus, the confusing matrix can described as quality indicator for the classification. For the fieldwork classification (Figure 2.10) a confusing matrix (Table C.1) has been created and is shown in Appendix C. The producer's accuracy defines the probability that the validation data has the same class as the classified data (Congalton, 1991). The producer's accuracy for a class is calculated in Equation 3.9, where e_{ii} is the correctly classified class and e_{i+} is the total number of observations in row i. Using data from the confusion matrix in Table C.1, the producer's accuracy for Grasslands is $\frac{392}{412} = 95.15\%$

$$Producer's\ Accuracy = \frac{e_{ii}}{e_{i+}} \tag{3.9}$$

The user accuracy is the probability that the classified location has the same class as the validation data (Congalton, 1991). The user accuracy is calculated in Equation 3.10, where e_{+i} the total number of observations in column i. Using the confusion matrix in Table C.1, the user accuracy for grasslands is $\frac{392}{616} = 63.64\%$.

$$User\ Accuracy\ = \frac{e_{ii}}{e_{+i}} \tag{3.10}$$

Besides the producer's and user accuracy, there are two accuracy indicators for the whole classified area: the overall accuracy and the kappa coefficient. The overall accuracy describes how many of the classified data locations, have been classified well using the validation data and vice versa (Congalton, 1991). The overall accuracy is calculated in (Equation 3.11), n denotes the total amount of entries in the matrix. For the confusion matrix in Table C.1 the overall accuracy is 58.14%.

$$Overall\ Accuracy\ = \frac{e_{ii}}{n} \tag{3.11}$$

Another classification quality indicator is the *kappa* coefficient, $\hat{\kappa}$. The $\hat{\kappa}$ is a measure of agreement. The closer the $\hat{\kappa}$ is to one, the more accurate the classification is (Congalton, 1991). The $\hat{\kappa}$ is calculated in Equation 3.12 and for the confusion matrix in Table C.1 the $\hat{\kappa}$ is 0.39.

$$\hat{\kappa} = \frac{n \times \sum_{i=1}^{r} e_{ii} - \sum_{i=1}^{r}(e_{i+} \cdot e_{+i})}{n^2 - \sum_{i=1}^{r}(e_{i+} \cdot e_{+i})} \tag{3.12}$$

4

Results

Within this chapter the classification maps, based on LiDAR-data, will be presented, the results from the accuracy indicators will be given. The LiDAR map will be compared to the Corine land cover classification (Figure 2.4) and to the fieldwork classification (Figure 2.11). At last, a discussion will be presented on using LiDAR and Sentinel-2 data.

4.1. Classification maps based on LiDAR

The Random Forest Classification map is presented in Figure 4.1. The Nearest Neighbour Classification map is presented in Figure 4.2. Besides, a boolean map has been made using the two classification maps and will be presented in Figure 4.3.

4.1.1. Random Forest

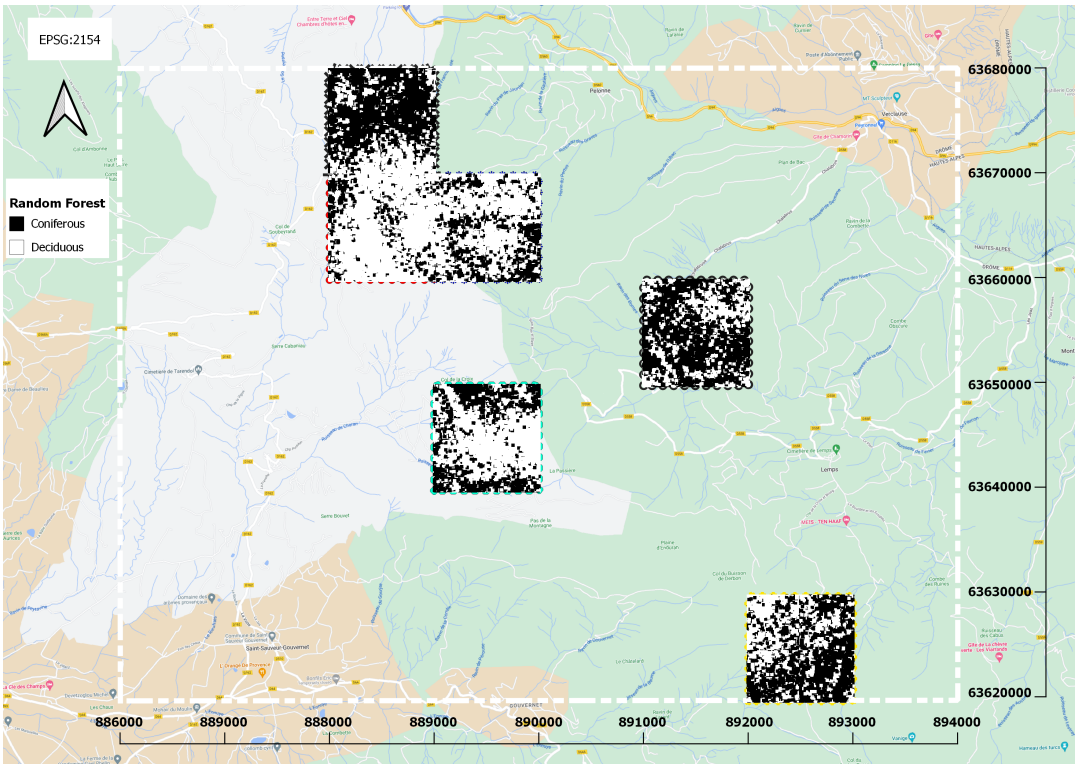


Figure 4.1: Classification map using the Random Forest Classification Algorithm in QGIS

The accuracy from the Random Forest classification method has been evaluated, where the training

data is compared to the validation data. The confusion matrix is produced and shown in Table 4.1. The accuracy of the Figure 4.1 is presented in Table 4.2. The user and producer's accuracy for the classes are over 92%. The overall accuracy is 93.05%. The $\hat{\kappa}$ has with a value of 0.86.

Table 4.1: Confusion Matrix for the Random Forest Classification Method

Random Forest	Coniferous	Deciduous	Total
Coniferous	2073	164	2237
Deciduous	133	1903	2036
Total	2206	2067	4273

Table 4.2: Accuracy results for Random Forest Classification

Random Forest	Producer's accuracy	User Accuracy
Coniferous	93.97%	92.67%
Deciduous	92.07%	93.47%

4.1.2. Nearest Neighbour

For the Nearest Neighbour, the following classification map has been created in Figure 4.2. The confusion matrix is created as well (Table 4.3), and thus the accuracy of the method is estimated in Table 4.4. The overall accuracy has a value of 81.89%. Compared to the accuracy estimation of Random Forest classification (Table 4.2), the producer's, user and overall accuracy for the Nearest Neighbour classification method are lower, with 85% as the maximum accuracy. Besides, the $\hat{\kappa}$ has a value of 0.64, which is less more accurate than the value of 0.86 calculated for the Random Forest classification (Table 4.2).

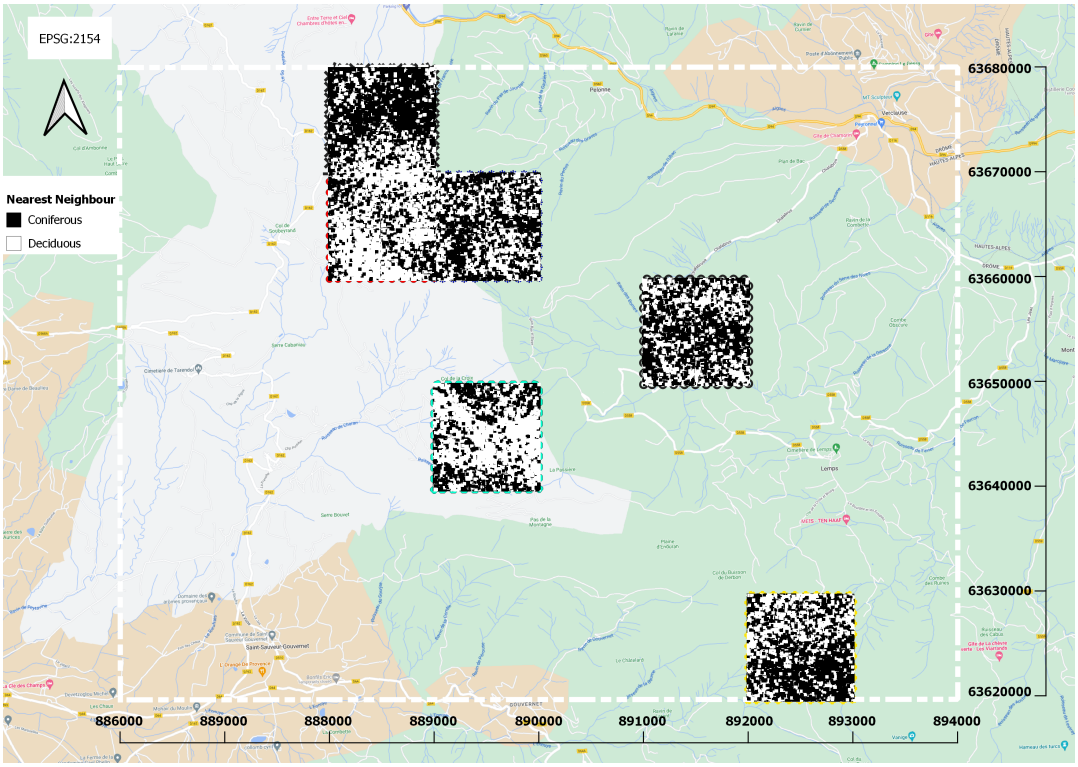


Figure 4.2: Classification map using the Nearest Neighbour Classification Algorithm

Table 4.3: Confusion Matrix for the Nearest Neighbour Classification Method

Nearest Neighbour	Coniferous	Deciduous	Total
Coniferous	1766	471	2237
Deciduous	303	1733	2036
Total	2069	2204	4273

Table 4.4: Accuracy results for Nearest Neighbour Classification

Nearest Neighbour	Producer's accuracy	User's accuracy
Coniferous	85.36%	78.95%
Deciduous	78.63%	85.12%

4.1.3. Comparing Random Forest and Nearest Neighbour

There are some differences between the Random Forest and Nearest Neighbour classification methods. A Boolean map is created for the two classification methods, the areas which are equal and those which are not are displayed in Figure 4.3.

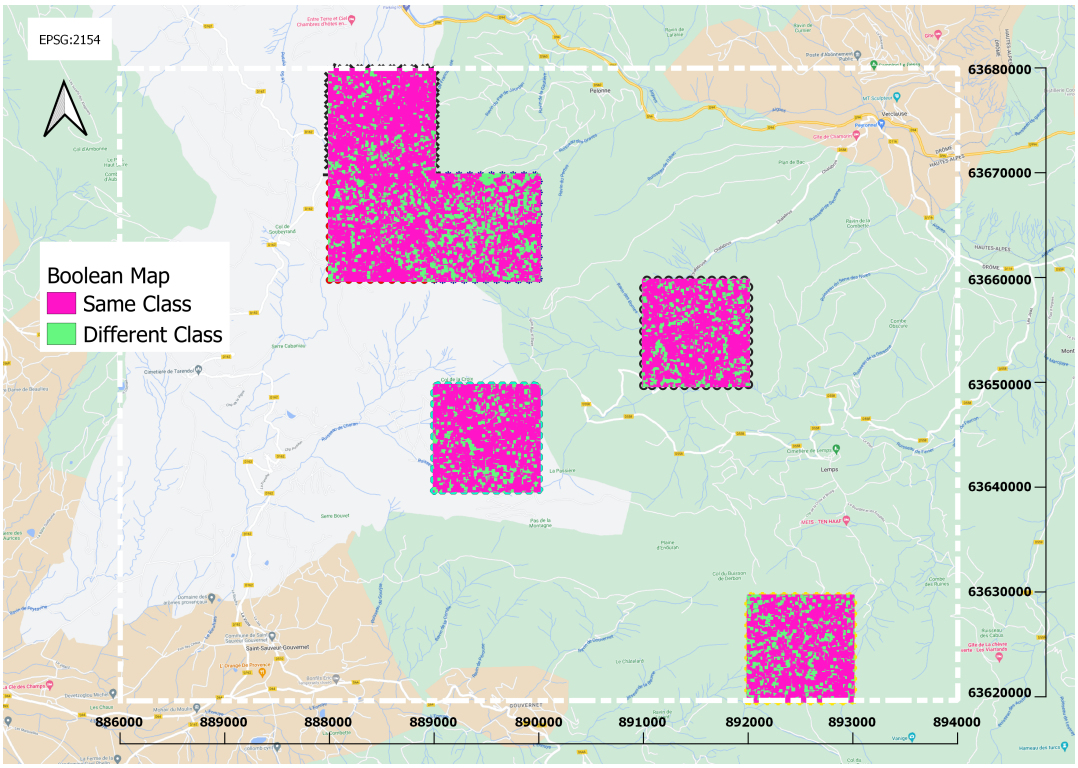


Figure 4.3: Boolean map from the Random Forest classification (Figure 4.1) and the Nearest Neighbour classification (Figure 4.2).

Using Google Maps Streetview a comparison will be made on the area to check the differences at a particular location. The location is visualised in the Appendix D. The comparison for the two maps is shown in (Figure 4.4), where the Google Street View picture above the two maps indicates looking North from the asterisk and the picture below the two maps indicates looking South. When looking North, some deciduous trees can be seen as well as coniferous trees. Whereas looking to the South mainly coniferous trees are visible. Looking at the two maps in Figure 4.4, the Random Forest Classification maps shows that there are some deciduous trees in the area, look for the blue dots. Whereas the Nearest Neighbour classification map does show less presence of deciduous trees (in white) looking

North with respect to the the asterisk.

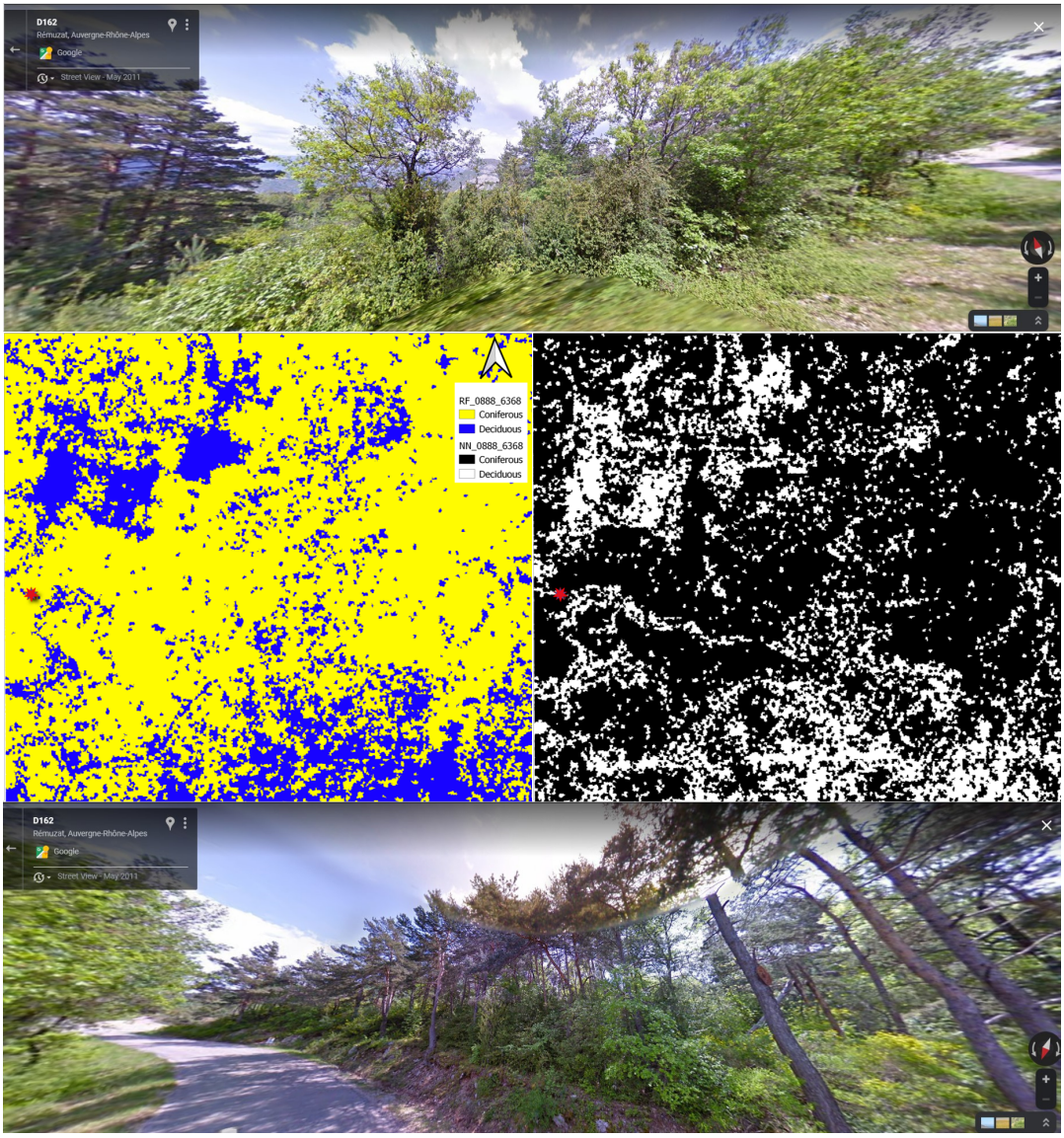


Figure 4.4: Zoomed-in comparison Random Forest (left and indicated the legend by RF) and Nearest Neighbour (right and indicated in the legend by NN) classification maps. Above the two maps a Google Street View visualisation is shown looking to the North with respect to the the asterisk, and below the map is looking to the south with respect to the asterisk.

The accuracy of the classification map using Random Forest (Figure 4.1) is higher than the accuracy of the Nearest Neighbour classification map (Figure 4.2). Random Forest Classification bases the classification on the decision trees. As could be seen in Figure 2.9, the decision trees are big and thus, the probability to classify it incorrectly decreases. Nearest Neighbour classification bases the decision on the neighbouring classes, it takes the value of the closest neighbours. Less calculations are made, and thus might lead to a lower accuracy. Thus, for the comparison with the Sentinel-2 derived land cover classification the Random Forest Classification maps Figure 4.1 are used.

4.2. Classification maps LiDAR compared to Sentinel-2

The classification map for Random Forest classification map Figure 4.1 will be compared to the Corine land cover classification (Figure 2.4) and the fieldwork 2021 classification (Figure 2.11).
Firstly, the map Figure 4.5 will be analysed, which is the Random Forest classification map (Figure 4.1)

overlaid with the Corine land cover map (Figure 2.4). The map shows similarities, Broad-leaved forests could be read as deciduous trees in the Figure 2.4. Mixed Forest is a mix between coniferous and deciduous trees. The coniferous trees are usually classified as coniferous trees and deciduous as deciduous trees. The mixed area, gives indeed a mixed classification of deciduous and coniferous trees.

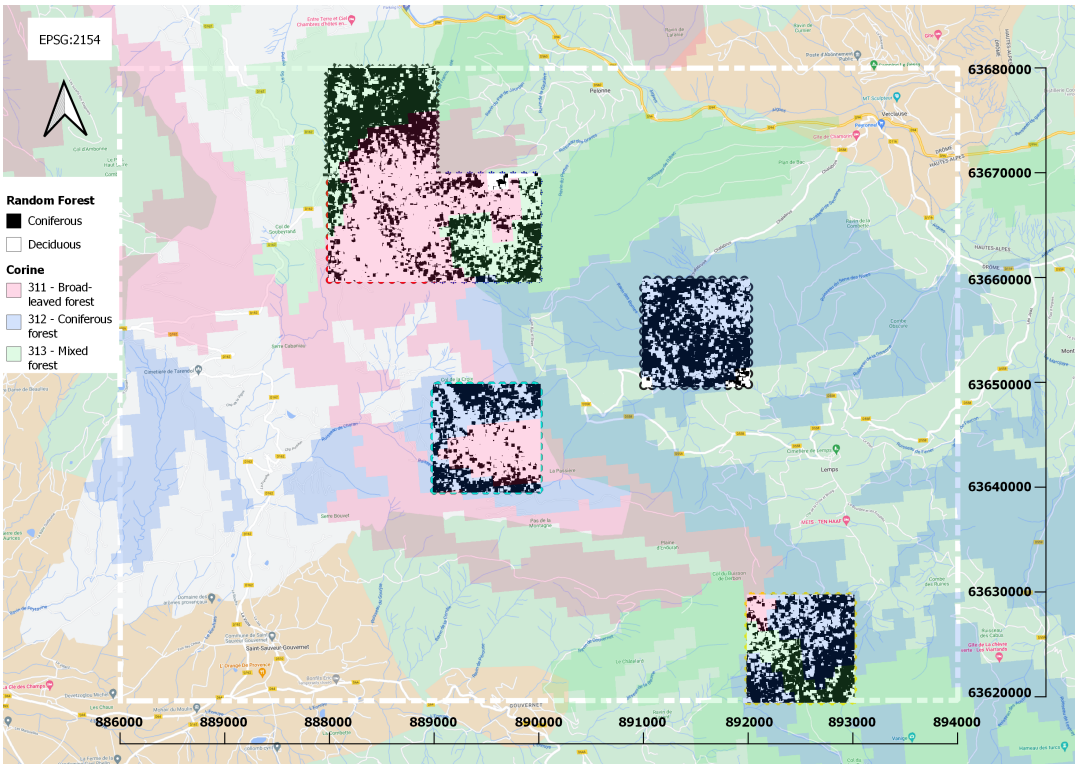


Figure 4.5: Random Forest classification map (Figure 4.1) overlaid with the Corine Land Cover Classification (Figure 2.4)

The Random Forest classification map (Figure 4.1) is overlaid with the 2021 fieldwork classification map (Figure 2.11). It gives the result presented in Figure 4.6. In comparison to Figure 4.5, this classification gives a result with a lower quality. There is lots of confusion between the classes, lots of coniferous trees are classed as deciduous trees as could be seen in the tile that ranges between [889000,890000,6364000,6365000]. Besides, the accuracy of the classification of Figure 2.10 is lower ($\hat{\kappa} = 0.39$) compared to 0.86 from the Random Forest classification (Figure 4.1).

4.3. Discussion

Several reasons for the differences between the classification for Random Forest, Corine and Fieldwork could be given. For the the fieldwork classification, 10 samples per class were obtained as training data. For the classification based on LiDAR 17093 patches were used, which means around 8,500 samples per class. From the training data, more data could be acquired about the features as more sampling could be done. It has to be said, that the training data for the LiDAR data was mainly acquired using the Corine Land Classification. Corine has a resolution of 100 metres. The map could contain, for instance, 60 metres deciduous trees and 40 metres coniferous trees, it will still be classified as deciduous. This could lead to falsely identifying the training data. Thus might lead to errors in classification. The resolution for the Random Forest Classification is five metres and the resolution for the Fieldwork classification is two and a half metres, those are determined much more precisely than the Corine Land Cover data.

Another discussion point that could lead to falsely classifying the area is the distinction in trees, the tree data used on the region is from 2005 to 2014 (Auvergne-Rhône-Alpes, 2019). The data said

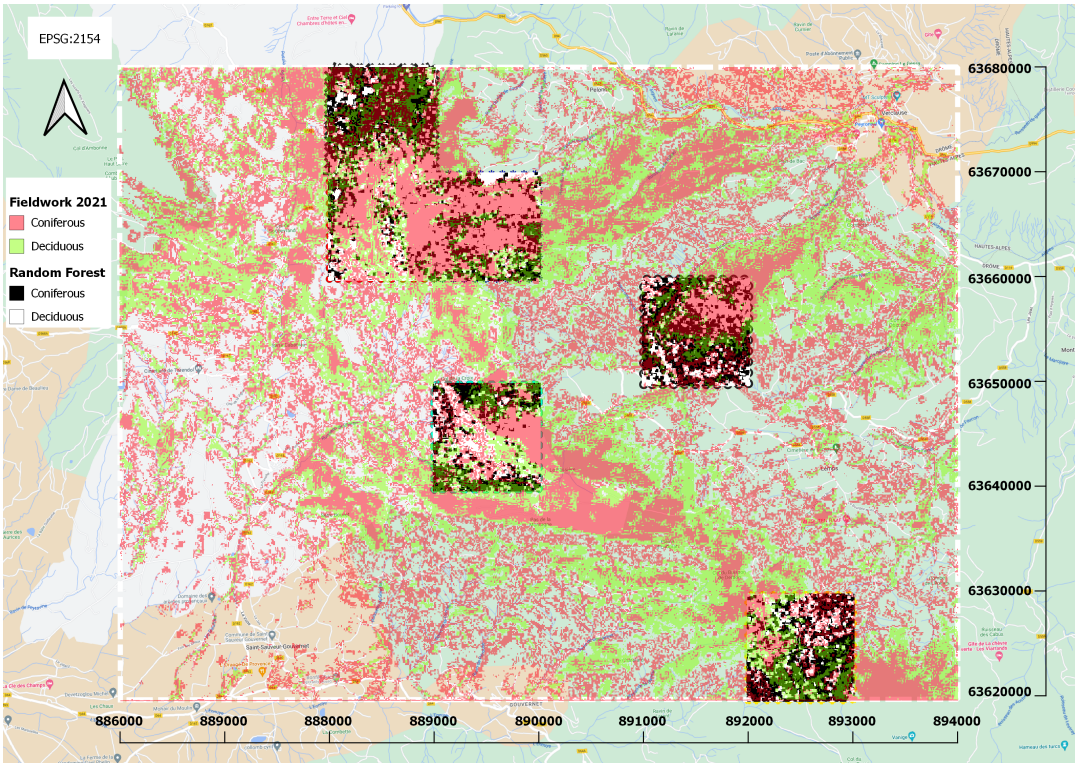


Figure 4.6: Random Forest classification map (Figure 4.1) overlaid with the Fieldwork 2021 classification for trees (Figure 2.11)

68.2% of the trees are the Pinus sylvestris, Abies Alba, Quercus Pubescens and Fagus sylvatica. In the meantime, the composition of trees could have changed by bushfires and logging. Besides, the coniferous trees Pinus sylvestris and Abies alba are both pine trees, and thus share similarities. Whereas for the deciduous trees, the Quercus pubescens is an oak and Fagus sylvatica a beech. It could be the case, that the selected deciduous trees are only collected for one tree species, and not both. Therefore there could be confusion within the training data for deciduous trees.

5

Conclusion

The purpose of this report was to analyse whether the available Airborne LiDAR data could enhance the land cover classification for the fieldwork region in France, the main research question of the project is *To which extent can LiDAR data be used to enhance the land cover classification for coniferous and deciduous trees?* To do so, the LiDAR data has been analysed and a classification map has been made. Using the LiDAR data could help to enhance the land cover classification from the fieldwork.

The LiDAR data could be extracted. Features have been derived from the data for coniferous and deciduous trees. The unique features can be used as classifiers for land cover classification. Firstly, the Random Forest classification map (Figure 4.1) has been made. The most important features for that classification could be derived (Table 3.3). Features based on the intensity were found to be very important. Using the important features the Nearest Neighbour classification method (Figure 4.2) was applied. The two acquired classification maps have been compared (subsection 4.1.3). Based on accuracy, an in-field comparison with Google Street View between the two maps has been conducted and a boolean map is analysed. Using the comparison, Random Forest returned a higher accuracy ($\hat{\kappa} = 0.86$) than the Nearest Neighbour classification method ($\hat{\kappa} = 0.64$). Random Forest classification showed the presence of deciduous trees at the in-field location (Figure 4.4), whereas Nearest Neighbour did not. From the boolean map (Figure 4.3), it could be seen that there was overlap in the classes.

Based on the in-field comparison and accuracy estimations, the Random Forest classification map has been compared with the Corine Land Cover Map (Figure 2.4) and the Fieldwork Classification (Figure 2.11). The Random Forest classification map shows lots of overlap with the Corine land cover classification (Figure 4.5), and less with the classification with the fieldwork classification (Figure 4.6). The map acquired during the fieldwork, had also a lower $\hat{\kappa}$ value of 0.39. This indicates a lower accuracy. For the fieldwork, ten samples of training data per class have been used instead of the 8,500 per class for the Random Forest classification. The training data for the Random Forest classification has been based on the Corine map. The Corine map has a resolution of 100 metres, which means that at least 51 metres within the 100 metres should contain of a class to be assigned to that class. Thus, the training data might be mixed with other classes. The deciduous trees in the region are both beeches and oaks, it is unknown if both beeches and oaks have been selected in the training data. Hence, some deciduous areas could be incorrectly classified.

For future research, the LiDAR data could enhance the classification for the fieldwork. The classification could be enhanced if more training data is used, if the training data is correctly selected and if it contains all tree species within the classes. Thus it can be trained better, best way would go into the field and take time to select training data. The same classification methods could be used. Some changes within the classification parameters could be applied such as selecting the amount of features in a decision tree and the amount of decision trees. Using the created classification map, it could enhance the accuracy of classification acquired during the fieldwork.

A

Tiles

Within the Appendix Tiles, the tiles used for the area are displayed. All tiles are retrieved from https://geoservices.ign.fr/lidarhd, and a *tile package* contains 4 1x1 kilometer tiles. For instance, the package *0886_6367* contains the following tiles: *0886_6367, 0886_6368, 0887_6367, 0887_6368* and the area that every tiles encompasses would be for example from the tile *0887_6366*, the tile would range from [886000,887000,6366000,6367000] within the coordinate system (EPSG:2154).

Table A.1: Tiles used in the project including retrieval date

Tile No.	Date Retrieved
0886_6367	25/11/2022
0886_6365	25/11/2022
0886_6363	25/11/2022
0886_6361	07/12/2022
0888_6367	25/11/2022
0888_6365	25/11/2022
0888_6363	25/11/2022
0888_6361	07/12/2022
0890_6367	25/11/2022
0890_6365	25/11/2022
0890_6363	25/11/2022
0890_6361	07/12/2022
0892_6367	25/11/2022
0892_6365	25/11/2022
0892_6363	25/11/2022
0892_6361	07/12/2022
0894_6367	25/11/2022
0894_6365	25/11/2022
0894_6363	25/11/2022
0894_6361	07/12/2022

B

Updated Features

The features in Table B.1 have been selected as there were clear differences. For the relative height there is almost $100cm$ of difference between the trees. The difference in standard deviation for the heights is too small. Kurtosis for the height are both negative. Skewness based on height for coniferous is positive, for deciduous is negative and thus the skewness is selected. The variance for the height has been included, this is calculated over the whole dataset instead of the relative height, as the variance in the coniferous height is more than 30% bigger than the deciduous variance. There was no big difference in the coefficient of variation for the height, thus it is not in the selection. The average values of the intensity show a big difference (1400 vs 1880), and therefore it is chosen. There is a big difference for the standard deviation of the intensity, the variance is the standard deviation squared. Thus, just the variance is selected. The kurtosis of the intensity is selected because it is mainly leptokurtic for coniferous trees, and platykurtic for deciduous trees. No big difference for the skewness and coefficient of variation for the intensity were measured and thus it is not selected. From the ratios, only the first ratio is selected as the returned values differed from each other. The lower part and upper part ratios are selected as well because for coniferous the lower part is bigger and the upper part smaller and for deciduous the other way around. The average Return Number and Number of Returns are included as well as it shows a difference. A fact about the data: even though the area of the coniferous ($8.01m^2$) is smaller than the deciduous ($8.61m^2$), on average it contains more points than the deciduous trees (540 vs 420). It seems to be denser.

Table B.1: Average values features for coniferous and deciduous trees, result for Table 3.2

features	avg_coniferous	avg_deciduous
rel_height_avg	847.27	749.67
rel_height_std	19.85	27.26
kurtosis_height	-0.72	-0.07
skewness_height	0.07	-0.09
var_height	86652.56	61485.40
CV_height	0.03	0.04
avg_intensity	1402.51	1879.86
std_intensity	778.80	1088.04
kurtosis_intensity	0.67	-0.54
skewness_intensity	0.81	0.39
var_intensity	637918.58	1213167.77
CV_intensity	0.56	0.61
ratio_first_part	0.35	0.29
ratio_second_part	0.17	0.19
ratio_third_part	0.24	0.26
ratio_fourth_part	0.24	0.26
lower_half_range_ratio	0.52	0.48
upper_half_range_ratio	0.48	0.52
len_point_data	539.67	420.01
area	80167.79	86149.76
avg_num_ret	1.77	1.66
avg_ret_num	2.51	2.29

C

Confusion Matrix Fieldwork

Table C.1: Confusion matrix for Classification fieldwork 2021 (Figure 2.10)

Spectral Angle Mapping Classification						**Reference Data**
Classification data	Grasslands	Farmed Trees	Urban	Coniferous Trees	Deciduous Trees	Total
Grasslands	392	4	10	6	0	412
Farmed Trees	6	661	283	25	0	975
Urban	2	29	202	0	0	233
Coniferous Trees	216	63	2	127	280	688
Deciduous Trees	0	0	0	404	29	433
Total	616	757	497	562	309	2741

D

Zoomed-in area for comparison

Within this appendix the zoomed-in area where comparison between the Nearest Neighbour and Random Forest classification takes place.

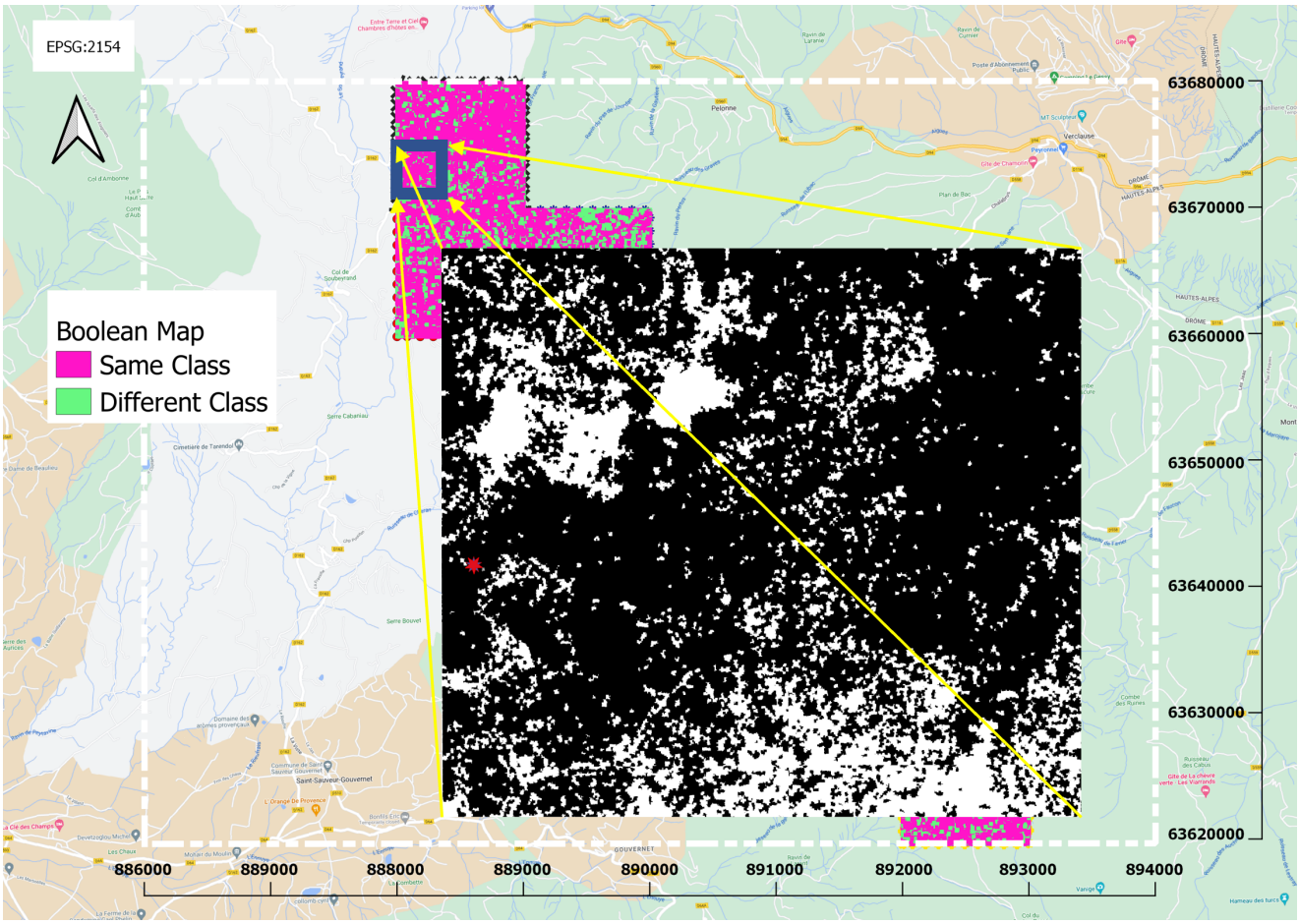


Figure D.1: The area where the comparison takes place with respect to the total fieldwork area (Figure 4.4). The comparison takes place within the tile *0888_6368*.

Bibliography

Abies alba. (2022). Retrieved December 17, 2022, from https://en.wikipedia.org/wiki/Abies_alba

ASPRS. (2008). https://www.asprs.org/wp-content/uploads/2010/12/asprs_las_format_v12.pdf

Auvergne-Rhône-Alpes. (2019). Prfb, Annexe-4 Massif Diois –Baronnies. https://draaf.auvergne-rhone-alpes.agriculture.gouv.fr/le-programme-regional-de-la-foret-et-du-bois-2019-2029-est-valide-a3112.html

Blom, J. C. (2021). Vesc fieldwork area division.

Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, *37*(1), 35–46. https://doi.org/10.1016/0034-4257(91)90048-B

Congedo, L. (2021). Semi-automatic classification plugin: A python tool for the download and processing of remote sensing images in qgis. *Journal of Open Source Software*, *6*(64), 3172. https://doi.org/10.21105/joss.03172

Corine land cover. (2018). Retrieved November 26, 2022, from https://land.copernicus.eu/pan-european/corine-land-cover

Dimitriadis, S., Liparas, D., & Tsolaki, M. N. (2018). Random forest feature selection, fusion and ensemble strategy: Combining multiple morphological mri measures to discriminate among healhy elderly, mci, cmci and alzheimer's disease patients: From the alzheimer's disease neuroimaging initiative (adni) database [A machine learning neuroimaging challenge for automated diagnosis of Alzheimer's disease]. *Journal of Neuroscience Methods*, *302*, 14–23. https://doi.org/10.1016/j.jneumeth.2017.12.010

Donges, N. (2019). *Intro to descriptive statistics*. Retrieved January 9, 2023, from https://towardsdatascience.com/intro-to-descriptive-statistics-252e9c464ac9

Durrant, T., de Rigo, D., & Caudullo, G. (2016). Fagus sylvatica in europe: Distribution, habitat, usage and threats.

ESA. (n.d.-a). *Multispectral instrument (msi) overview*. Retrieved November 30, 2022, from https://sentinel.esa.int/web/sentinel/technical-guides/sentinel-2-msi/msi-instrument

ESA. (n.d.-b). *Sentinel-2*. Retrieved November 19, 2022, from https://sentinel.esa.int/web/sentinel/missions/sentinel-2/instrument-payload/resolution-and-swath

Fagus sylvatica. (2022). Retrieved December 13, 2022, from https://en.wikipedia.org/wiki/Fagus_sylvatica

Farjon, A. (2017). *A handbook of the world's conifers*. Brill. https://books.google.nl/books?id=IXchMQAACAAJ

Heywood, J. (2020). *Land type classification of icesat-2 global geolocated photon data* (Doctoral dissertation).

Labaar, A. L. (2022). *Dune vegetation classification using uav-lidar point clouds* (Doctoral dissertation).

Li, J., Lin, S., Yu, K., & Guo, G. (2021). Quantum k-nearest neighbor classification algorithm based on hamming distance. *Quantum Information Processing*, *21*(18). https://doi.org/10.1007/s11128-021-03361-0

Lindenbergh, R. C. (2021). Lecture 4: Deterministic interpolation and lidar.

Luo, S., Miao, D., Zhang, Z., & Wei, Z. (2020). Non-numerical nearest neighbor classifiers with value-object hierarchical embedding. *Expert Systems with Applications*, *150*, 113206. https://doi.org/doi.org/10.1016/j.eswa.2020.113206

Optics, E. (n.d.). *Light source measurement*. Retrieved November 26, 2022, from https://www.eckop.com/applications/light-source-measurement/

Pinus sylvestris. (2022). Retrieved December 16, 2022, from https://en.wikipedia.org/wiki/Pinus_sylvestris

Quercus pubescens. (2022). Retrieved December 14, 2022, from https://en.wikipedia.org/wiki/Quercus_pubescens

Rushforth, K. (1999). *Trees of britain and europe*. HarperCollins. https://books.google.nl/books?id=hBh0QgAACAAJ

Sokal, R. R. (1974). Classification: Purposes, principles, progress, prospects. *Science*, *185*(4157), 1115–1123. https://doi.org/10.1126/science.185.4157.1115

Vosselman, G., & Maas, H.-G. (2010). Whittles Publishing. https://app.knovel.com/hotlink/toc/id:kpATLS0002/airborne-terrestrial/airborne-terrestrial

Zbigniew, W. (2017). Mp estimation applied to platykurtic sets of geodetic observations. *Geodesy and Cartography*, *66*(1), 117–135. https://doi.org/10.1515/geocart-2017-0001

Zwillinger, D., & Kokoska, S. (2000). *Crc standard probability and statistics tables and formulae*. Chapman & Hall, New York.