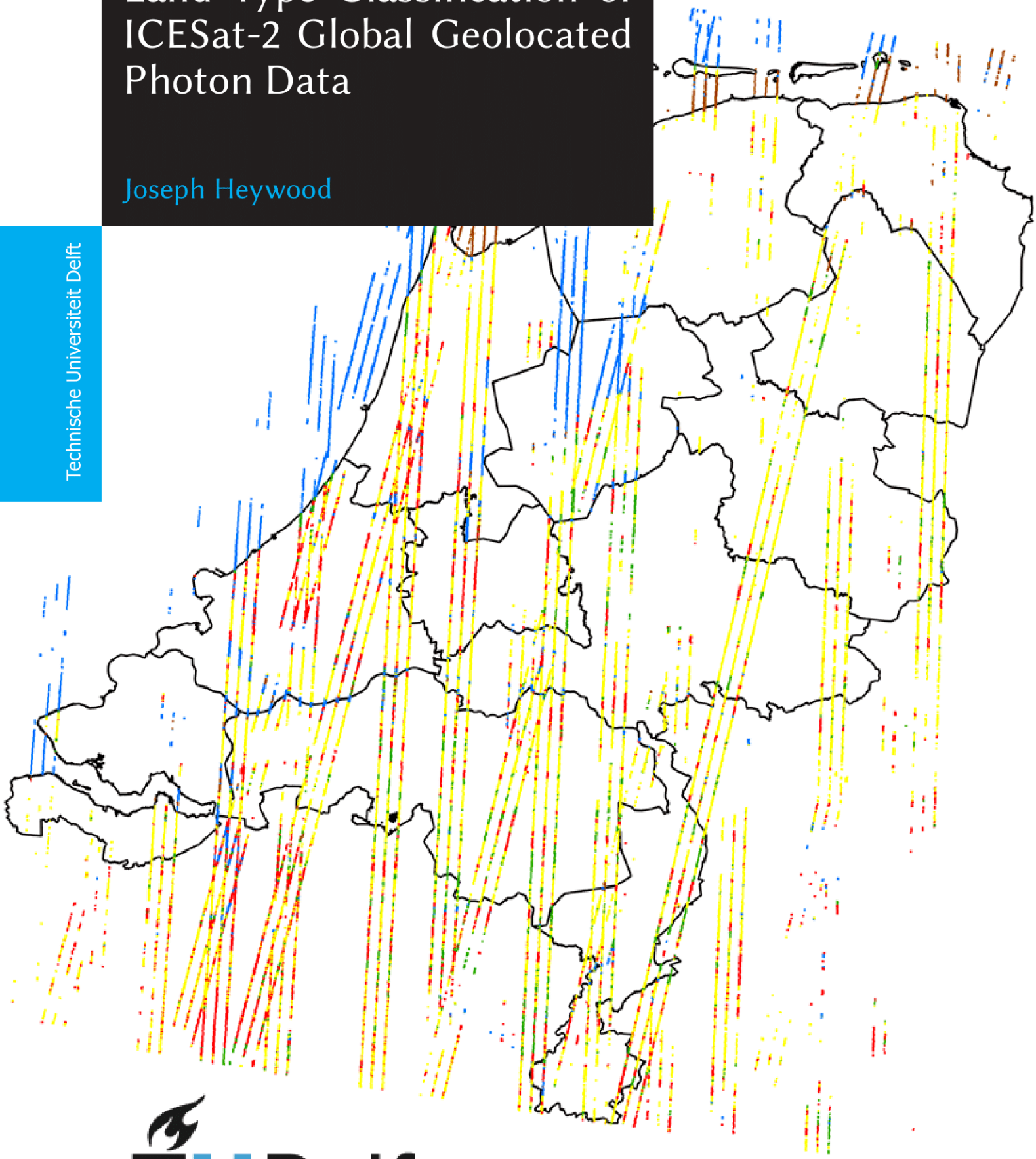
Land Type Classification of ICESat-2 Global Geolocated Photon Data

Joseph Heywood

Technische Universiteit Delft



TU Delft

Challenge the future

Land Type Classification of ICESat-2 Global Geolocated Photon Data

A thesis report presented for the degree of Master of Science
at

Delft University of Technology

Faculty of Aerospace Engineering
Faculty of Civil Engineering and Geosciences

September 10, 2020

by

Joseph Heywood 4219554

An electronic version of this thesis is available at https://repository.tudelft.nl/

Supervisors:

Dr. Roderik Lindenbergh Department of Geosciences & Remote Sensing
Dr. Wouter van der Wal Department of Space Engineering

To Peter,
We're gonna miss you,
I hope you're at peace now, mate.

Preface

The delivery of this report marks a monumental moment for me. It's been 7 years since I was sat in a café in Thailand and decided I was going to give it a go at applying to a couple of uni's on the mainland. Little did I know what was to come. A couple of exams, a handful of cities, a trip to Australia and countless experiences later, it has come to the end.

It goes without saying that the enjoyment and motivation I have experienced for the last 7 years would not have been possible without the absolute legends I am lucky enough to have in my life. Lucas, who was, and has been there from the start of all of this. The Oostplein boiz, Thomas, Tihab and Mike, who have always so beautifully put me in my place, r3kd. The JVB crew, Sietsfge, SchotmanG, Mique, Nox and Wout, who in sharing a home with, have now made The Netherlands a home to me. The beys'n'cigs crew, who always bring me back to my childhood. More recently, Robbo, your selfless support during the rollercoaster of stress that comes with a thesis is much appreciated.

I'm sure that my family thinks I have been dossing for the last 7 years. Nevertheless, you continue to support everything I undertake. Mum, my capacity to pass this degree comes from you. Dad, your stability and advice on all my important decision is invaluable to me. Grandmas and Grandads, your unwavering belief in everything I do continues to give me confidence in myself.

Finally, I would like to thank Roderik and Wouter. I honestly couldn't have wished for more supportive, insightful and pleasant supervisors. After every meeting with you both, I came out feeling confident about my work and motivated to keep going. I am genuinely pleased with this report and I hope you are too.

Abstract

NASA's ICE, Cloud and Land Elevation Satellite-2 (ICESat-2) has been measuring the topography of the Earth's surface since its launch in September 2018. Equipped with a single instrument, namely, the Advanced Topographic Laser Altimeter System (ATLAS), ICESat-2 is able to acquire the along track vertical profiles of its laser footprints. While satellite based land classification has traditionally been performed with the use of multi-spectral data, which doesn't consider the vertical structure of the surface in question, the three-dimensional Light Detection and Ranging (LiDAR) product provided by ATLAS allows for the observation of the vertical structure of the illuminated surface. This provides information for the discrimination of surfaces that are only distinctly different from one another in this dimension, such as different types of vegetative species. Moreover, a greater understanding of how the signal interacts with different land types will benefit current and future users of the data. This study presents a first look at the potential of the base scientific data set provided by ATLAS, "ATL03", as a means of land type classification. Features extracted from ATL03 vertical profiles are used to classify multiple land types in The Netherlands, namely, "Artificial Surfaces", "Agricultural Areas", "Forest and Semi-Natural Areas", "Wetlands" and "Water Bodies". $100m$ grid cells were classified and validated with the CORINE land cover database. The overall classification accuracy was 71.2%, however, after a visual inspection of the misclassification errors it was found that that the actual accuracy was a minimum of 5.5% higher, that is, 76.7%. 51 features were created to discriminate between land classes and their importance per class was analysed. In general, simple statistical parameters, such as the standard deviation and percentile ranges worked well in distinguishing between classes. For the classes with a greater vertical range, such as "Artificial Surfaces", features that described the height and prominence of its scattering surfaces were most important.

Contents

List of Figures

List of Tables

1. Introduction

Land cover maps provide segmented information pertaining to the different types of physical coverage on the Earth's surface, such as Forests, Wetlands, Artificial surfaces, Agriculture and Water Bodies. The classification of land cover is an essential component of socio-economic policy and decision making in the development, planning and management of critical infrastructure and resources. Land cover assessment and the monitoring of its dynamics are essential requirements for the sustainable management of natural resources and environmental protection. In addition, characterisation of the landscape is also necessary to establish boundary conditions for models which are sensitive to these properties, such as predictive models of atmospheric change that depend on land-atmosphere interactions. The most widely used remote sensing technique for land classification is the visual and computational interpretation of high resolution single and multi-spectral satellite imagery. The spectral data provided by remote sensing satellites such as Landsat and Sentinel have allowed the mapping of land cover on a national and continental scale [Feranec et al., 2016][NERC Environmental Information Data Centre, 2019][Hazeu, 2013].

The Ice, Cloud and Land Elevation Satellite-2 (ICESat-2) was launched in September 2018 and carries a single instrument, the Advanced Topographic Laser Altimeter System (ATLAS). It is the successor to the original ICESat mission, which similarly carried a single laser altimeter. ATLAS sends out 3 pairs of laser beams that reflect off the Earth's surface and records the return travel time of individual photons. By measuring the along track return of individual photons, ATLAS is able to trace the vertical profile of the landscape, providing a slice of the topographical variation at the location of the footprints. This offers a unique perspective of observation in comparison to spectral data as it provides the opportunity to obtain information pertaining to the vertical structure of the illuminated surface. This can be particularly usefull for surfaces that are only distinctly different from one another in the vertical dimension, such as different types of vegetative species, the identification of wildland fuel or the distinction between woodland and agricultural areas.

The main scientific goals of the ICESat-2 mission are to measure a changing cryospheric environment in order to understand the processes that connect the polar regions with the global climate system. However, secondary goals such as the worldwide measurement of canopy height to measure carbon storage provide insight into the wider possible applications of the data. In addition to the official scientific goals associated with the mission, NASA's "Early Adoption" programme promotes research applications into a wide range of fields in order to provide a fundamental understanding of ICESat-2's raw signal [NASA, 2019]. Many potential use cases of ATLAS data are proposed, such as ice forecasting, wildland fuel assessments and coastal mapping/monitoring. Similarly to the wide range of proposed applications in the early adopter programme, it is intended that the research provided in this report will demonstrate the utility of this novel data product outside the predominant scientific goals.

The aim of this research is to classify land types in The Netherlands with the use of the vertical profiles collected by ATLAS. Specifically, with the use of the base scientific dataset, "ATL03", which provides the latitude, longitude and vertical heights for each reflected photon along the laser footprints of ATLAS. The data products collected by ICESat-2 are the first Geo-referenced photon data product available on a world scale and by analysing the behaviour of ATLAS data when interacting with different surfaces, it is hoped that a greater understanding of the capabilities and limitations of the data set will be acquired. A greater understanding of the behaviour of the signal with the land types analysed in this research can be beneficial to current and future users of ICESat-2 data. The study is intended as a means of understanding which unique features to ICESat-2 are the most important for classifying land and what these features can reveal about the surface under illumination.

While this research is the first example of multi-class land type classification of ATLAS data, there has been some research suggesting its potential. [Duong, 2010] and [Molijn et al., 2011] showed that the more rudimental footprints recorded on the original ICESat mission were able to successfully classify high level land types such as "urban areas" and "vegetation". In both cases, the full waveform was simplified

into a number of Gaussian modes, from which characteristic attributes were extracted for classification. [Zhou et al., 2016] argued that by discretisizing the full waveform, one is not taking full advantage of the information in the waveform. The original ICESat mission classified elliptical laser pulse footprints of $\approx 70m$ in diameter every $175m$. In contrast, ICESat-2 collects a continuous stream of individual photons along its footprints, collecting a significantly higher amount of information in the vertical topographical plain.

There is research that shows the potential of using ATLAS profiles for land classification. [Gwenzi et al., 2016] showed that the histograms made from the ATLAS profiles provide characteristic attributes that show vegetation is present as well as the height and other attributes of the vegetation canopy. Furthermore, in [Neuenschwander and Magruder, 2019] it can be seen that different types of vegetation provide observable differences in profiles. For example, the standing water below a mangrove forest was clearly identifiable from a visual inspection of ATLAS profiles. It is expected that the information rich profiles will be able to successfully detect features attributed to other land types found in The Netherlands, such as man-made structures, agriculture and rivers. In fact, [Zhang et al., 2020] performed a successful binary classification of ice/snow covered and bare land with only the returned photons that were considered noise. It has been shown that the original ICESat mission was able to successfully classify high level land types and that ICESat-2 can identify distinguishable features in its vertical profiles. However, it has not been attempted to classify areas of dense land type diversity, such as The Netherlands. In addition, the behaviour of the ICESat-2 signal on surfaces within The Netherlands, such as Agriculture and Artificial Surfaces, is unknown.

The goal is to determine to what extent and how accurately ICESat-2 data can be used to classify land types. In doing so, it will be necessary to determine what attributes of the surface topography affect the return signal as well as which features of the ATLAS data product are most important at classifying certain land types. Features are henceforth defined as a measurable property or characteristic of a group of photons, such as the average height, range height or standard deviation of the height. As this classification serves as a baseline for the data product, it was decided to use the highest level land types provided by the validation data. These were "Artificial Surfaces", "Agricultural Areas", "Forests and Semi-natural Areas", "Wetlands" and "Water Bodies". In addition, due to the unique measurement technique which records a continuous stream of individual photons, a custom validation technique will be necessary.

The objectives of the research can be summed up by the following research question and sub-questions:

To what extent can ICESat-2 geo-referenced photon data be used to classify land cover types in The Netherlands?

1. *How effectively can ICESat-2's ATL03 global geolocated photon data be used to classify between the land types "Artificial Surfaces", "Agricultural Areas", "Forest and Semi-natural area", "Wetlands" and "Water Bodies" using the random forest machine learning technique?*

2. *Which features unique to ATL03 geolocated photon data are best at distinguishing land types from one another?*

3. *What technique can be used to validate the classification?*

The structure of the report is as follows. Section 2 will provide an introduction to the use of ICESat and ICESat-2 data in land classification as well as more detailed discussion about the ATLAS instrument and the Random Forest classification technique. The research methodology is provided in Section 3. Section 3.1 provides the reader with a first look and description of the ATLAS data, while the validation data and study area is provided in Sections 3.2 and 3.3 respectively. The classification strategy is discussed in Sections 3.4 and 3.5, with the test set-up presented in Section 3.6. The results are separated into 3 sub-sections. The overall accuracy, feature importance and a misclassification analysis can be found in Sections 4.1, 4.2 and 4.3 respectively. Finally, the conclusions and recommendations are presented in Section 5.

2. Background

The aim of this Section is to provide the reader with some relevant background information on the research topic. It is not necessary to read the information provided here for the understanding of the research, but it does serve as a sound basis for the readers unfamiliar with the ICESat missions and/or the working principle of ATLAS and/or the random forest classification algorithm.

The original ICEsat mission is introduced in Section 2.1 and ICESat-2 in Section 2.2. The working principle of ATLAS is discussed in Section 2.2.1 and an overview of the relevant research performed with ATLAS data is provided in Section 2.2.2. Section 2.3 provides an introduction to the Random Forest machine learning technique and associated parameters relevant for this research.

2.1 ICESat

ICESat, the predecessor to ICESat-2, was launched in 2003 and remained active until 2010. The original ICESat mission carried a single laser altimeter, the Geoscience Laser Altimeter System (GLAS) and its success led directly to the follow up mission that is ICESat-2. The main scientific objectives of ICESat-1 were to monitor the mass balance of the polar ice sheets and their contributions to global sea level change. Secondary objectives were to measure cloud heights and the vertical structure of clouds and aerosols. As oppose to ATLAS, which measures individual photons, GLAS measured the returns from pulses of energy fired toward the surface. Portions of this energy are reflected (and absorbed) at different points on the way to the surface, such as in a cloud or the canopy of a tree. The return signal is therefore expressed as the return power as a function of time and the waveform provides information on the vertical structure of the surface.

[Duong, 2010] performed a case study over The Netherlands in which it was assumed that the return waveform can be decomposed into different Gaussian components, where each component can describe the properties of a particular reflecting object within the footprint. From the decomposed Gaussian components, [Duong, 2010] derived some individual waveform parameters to be used in the land type classification: Number of modes, Total Energy, Waveform begin and Waveform extent. A graphical description of these parameters can be seen in Figure 2.1.

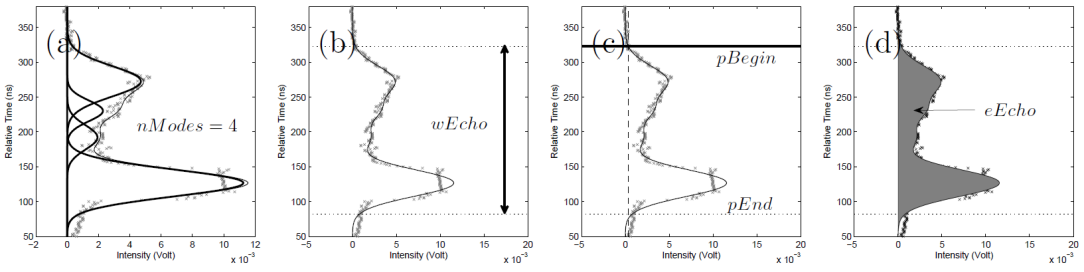


Figure 2.1: ICESat Waveform Classification Parameters derived from an arbitrary waveform: (a) Number of Modes, (b) Total Return Energy, (c) Waveform Begin, (d), Waveform Extent [Duong, 2010]

[Duong, 2010] used a simple decision tree classification scheme to discriminate between 4 different land types, that is, **high vegetation**, **urban**, **bare land/low vegetation** and **water**. The classification scheme simply used thresholds on the parameters to decide on the land type. For example, as water absorbs considerably more energy from the GLAS pulse than the other land types, it was decided that any return waveform with an echo energy of less that $45fJ$ would be considered water. The other land types were classified in the same way, making use of the rest of the classification parameters. The overall accuracy was 73%, indicating that the vertical structure of waveforms can be used to successfully classify land.

[Molijn et al., 2011] employed a similar technique in the Dry Valleys of Antarctica to different types of terrain, namely, **snow**, **rock**, **ice** and **water**. In a similar fashion to [Duong, 2010], [Molijn et al., 2011] felt that only 4 attributes were necessary as an input to the classification algorithm. A decision tree was employed in this study, which is a relatively simple classification scheme allowing the user to obtain direct insight into the influence of the different decisions on the classification outcome. The attributes used in [Molijn et al., 2011] were different to those used in [Duong, 2010] and it is clear from the discussion of the respective authors that the attributes needed for successful classification are dependant on the land types to be classified. As in [Duong, 2010], the overall classification accuracy of 74% indicated that the method can adequately classify these land types. Although, the ice and snow land types performed relatively poorly due to the fact that the attributes such as reflectivity were not able to discriminate between these two surfaces. With the heights of each photon returned along the footprint surface, ICESat-2 offers the ability to employ additional attributes that could better discriminate between these two classes.

[Zhou et al., 2016] thought that by discretisizing and simplifying the full waveform into a number of Gaussian modes and/or by only taking shape related metrics (echo width, skewness, kurtosis etc) from the waveform, that one is not taking full advantage of all the information present in the waveform. [Zhou et al., 2016] employed a curve matching approach in an attempt to access more of the information from the waveform. The curve matching approach involved the Kolmogorov-Smirnov (KS) test, which is a statistical measure of the equality of two probability distributions. The samples were classified by measuring the similarity in distribution between a test sample and a reference curve of a desired land class. [Zhou et al., 2016] argues that by using the full culmulative distribution function, one can extract the information provided by the Gaussian discretization as well as extra information contained in local peaks, skewness and kurtosis. This argument was somewhat validated in the results of the study where [Zhou et al., 2016] compared the rule based approach employed by [Duong, 2010] and [Molijn et al., 2011], to the proposed curve matching approach. The curve matching approach had a superior accuracy of 87.2% compared to 83.7% achieved by the rule based analysis. While these results should be taken with caution due to the different land types classified as well as different attributes employed, there is no doubt promise in the land classification of ATLAS data, given that there will be much more accessible information contained in its profiles, in comparison to those of the original ICESat mission.

2.2 ICESat-2

The second Ice, Cloud and Land Elevation (ICESat-2) satellite was launched on the 15th September 2018 as the successor to the original ICESat mission, which was operational from 2003 to 2009. The ICESat-2 mission allows the continuation of altimetric laser measurements of the Earths' polar regions that started in 2003 with the Geoscience Laser Altimeter System (GLAS) on-board the first ICEsat satellite. The data gap between these two satellites was partially filled with NASAs' "Operation IceBridge", an airbourne campaign using a variety of instruments to observe Arctic and Antarctic ice sheets, ice shelves and sea ice [NASA, 2018]. These three missions share a main scientific goal of observing the polar regions in unprecedented detail in order to better understand processes that connect the polar regions with the global climate system.

While the main scientific goal of the ICESat-2 mission is the observation of the Earths' polar regions [Markus et al., 2017], there are a host of other practical and scientifically interesting use cases. The wide variety of use cases is best described by the applicants of NASAs early adoption programme. This programme promotes ICESat-2 data research in a wide range of fields so that a fundamental understanding of potential use cases can be realised and scaled into organisations' policy, business, and management activities [NASA, 2019]. Current applicants have proposed research looking into, to name a few, wild land fuel assessments, quantifying biomass carbon sequestration, water level monitoring of lakes and reservoirs, volcanic hazard mitigation and river hydrodynamics estimation. The diverse set of use cases drawn from satellite laser altimetry is also exemplified by the successful studies that resulted from the original ICESat mission such as the resolution of mass balances of mountain glaciers and ice caps [Pritchard et al., 2009] [Kropáček et al., 2014], sea ice freeboard, thickness and volume measurements [Kwok et al., 2007] [Kurtz and Markus, 2012] and global vegetation height calculations [Lefsky, 2010]

[Simard et al., 2011].

2.2.1 ATLAS

The sole instrument on-board ICESat-2, the Advanced Topographic Laser Altimeter System (ATLAS) is a photon counting light imaging, detection, and ranging (LiDAR) system. Photons are sent from ATLAS to the ground and the time taken for them to return is recorded. With a knowledge of the precise orbit and attitude of the satellite, one can accurately determine the height of the spots on Earth illuminated by the satellite. The precise information of the orbit is determined by the ancillary systems such as the Global Positioning System (GPS) and star trackers. The measurement principle of ATLAS is fundamentally different to the GLAS instrument on-board ICESat which measured the full waveform return as opposed to the individual photons measured by ATLAS. ATLAS transmits green light laser pulses at $532nm$ at $10kHz$ (GLAS operated at $40Hz$). About 20 trillion photons leave the ATLAS instrument every pulse, with only about a dozen returning for measurement. The users of ATLAS data will be provided with "photon cloud" graphs, showing thousands of data points of the geo-referenced photons and the time it took them to return to the instruments. By applying further algorithms to remove noise and detect the underlying signal, one can determine the elevation of the various surfaces that will appear under ICESat-2 during her orbits.

ICESat-2 will transmit a total of 6 beams organised in a 2 x 3 array, where pair of beams will be separated by approximately $3.3km$ in the cross track direction. The distance between the lasers in a pair is approximately $90m$ and can be adjusted by changing the yaw angle of the satellite. The ATLAS beam and footprint geometry can be seen in Figure 2.2. The pairs of beams are each characterised by a strong beam and a weak beam, these can be seen as the light green and dark green beams in Figure 2.2, there is an energy ratio between the two beams of about 1:4. The footprint of each individual laser is approximately $14m$ wide, that is, photons can be returned to the instrument from within a $14m$ along track corridor.

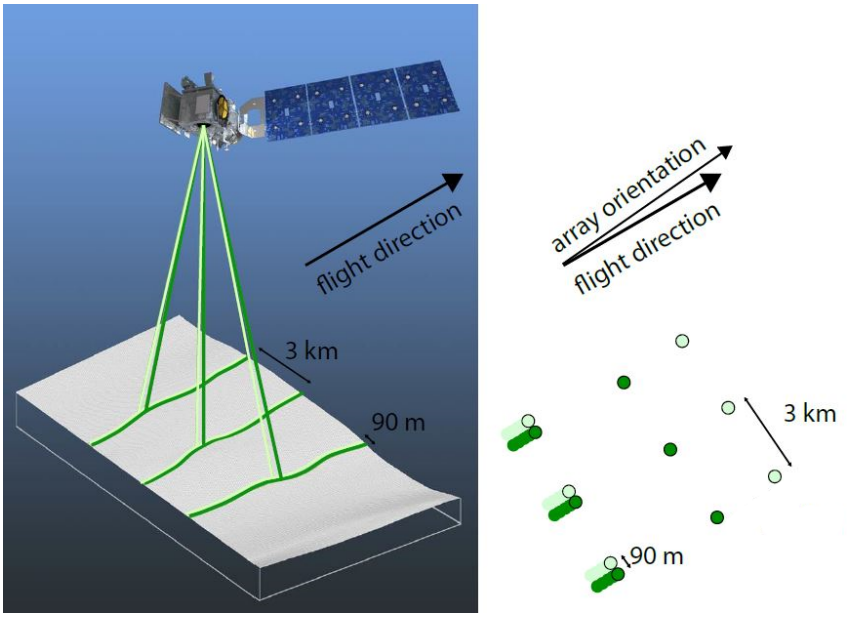


Figure 2.2: ATLAS Idealised Beam and Footprint Pattern [Neumann et al., 2018] (Left) ICESat-2 beam pattern (Right) Instantaneous footprint pattern

Assuming that the atmosphere is clear enough for photons to reach the desired surface, they are scattered upon arrival once or many times by the various land types they come into contact with. Some of these photons are scattered back towards the ATLAS receiver, which records them as a "photon event". Based on performance models, the arrival times of up to 12 photons per transmitted laser pulse can be

recorded by the detector electronics. In addition to the potential 12 signal photons that could be received [Neumann et al., 2018], noise photons from the sun are constantly entering the telescope. The on-board processing system knows an approximate height range that the photon should return from, so it can already ignore many of the photons that are returning from high in the atmosphere. However, photons are also arriving from the surface that have originated from the sun. Given that a photon is reflected from within the desired height range, all photons, regardless of origin, are time tagged and saved as what is referred to as a "photon event".

One of the primary tasks of the on-board data processing system is to determine which of the photon events have indeed been reflected from the laser pulse and which are noise photons entering from some other place. The workings of the algorithm itself are beyond the scope of this report and can be found in [Neumann et al., 2018], Section 5. The algorithm classifies each photon depending on the likelihood that it is a signal photon. It can be classified as "High", "Medium", "Low" and "Noise". For the research presented in this report, only the photons with a "high" classification are considered. It was decided to only classify "high" signal photons as it is desired to obtain the baseline performance of the dataset. It may be that the lower confidence photons make the underlying surface signal stronger, however, it could also have the opposite effect. The performance of the different photon signal confidences is out of the scope of this report.

2.2.2 ICESat-2 for Land Classification

As the ICESat-2 mission is largely a continuation of the original ICESat mission, the main principle of the LIDAR instrument is the same, but of course, with some functional improvements. ATLAS will operate three pairs of profiling lasers in comparison to the single beam operating on GLAS. This will allow for the determination of local cross-track slope and an improved spatial coverage.

Research has been conducted for various scientific objectives on individual land types such as high and low vegetation and land and sea ice. This will give insights into how the ICESat-2 signal behaves when classifying multiple land types. This will not only contribute toward the interpretation of the classification results but will also provide ideas as to what features/attributes should be used for specific land types as well as patterns that may appear in the data that are specific to a certain land type. Additionally, there are various further products that have been created from ATL03 data that provide height information for certain land types. In some cases, classification algorithms have been created in order to detect a certain land type or feature. These methods and their resulting products can be used as an aid in the classification algorithm.

One of the main scientific objectives is to measure the canopy height of vegetation around the world [Markus et al., 2017]. Naturally, to achieve this goal, one would need the means to identify and classify both high and low vegetation, such that further processing is performed in areas of interest only, that is, the calculation of canopy height is only performed at land types classified as vegetation. Authors have indirectly proposed methods to identify vegetation as a land class over a variety of ecosystems by showing the varied and distinct properties at these locations. This provides useful insights into the potential for land classification. For example, Figure 2.3 shows the diagrammatic representation of a canopy height calculation algorithm used to calculate tree height in a savanna ecosystem. As can be seen, the shape of the histogram indicates that there is a tree present and can provide an estimate of the height of the tree. By creating numerical parameter/s from the properties of the histogram, one could use this information as an attribute in the classification algorithm to indicate whether or not vegetation is present, the density of the vegetation and the height of the canopy.

Similarly, [Neuenschwander and Magruder, 2019] calculate the canopy heights in different ecosystems, such as tropical forests, savanna/woodland vegetation and mangrove forests. Figure 2.4 shows the ICESat-2 profile of a mangrove forest in Mexico. One can see both the similarities and differences between the savanna profile presented in Figure 2.3. The top of the canopy provides a similar curving profile as the savanna profile and other ecosystems presented in [Neuenschwander and Magruder, 2019], which will presumably be shared with all high vegetation land classes. Conversely, Figure 2.4 contains the specular returns which characterise this specific kind of vegetation, namely a mangrove forest. In

some cases, these specular returns can be used to identify standing water or saturated soils. Again, this kind of pattern could be recognised in a classification algorithm for more specific/lower level land types. More recently, [Liu et al., 2020] has taken to classifying even lower level vegetation types with a feasibility study in the mapping of areas of burnt vegetation. The unique vertical structure measured by ATLAS allowed the successful classification along ICESat-2's tracks achieving an accuracy of 83% with the use of random forest classification. [Liu et al., 2020] used 24 simple metrics such as average, medium and standard deviation of canopy height.

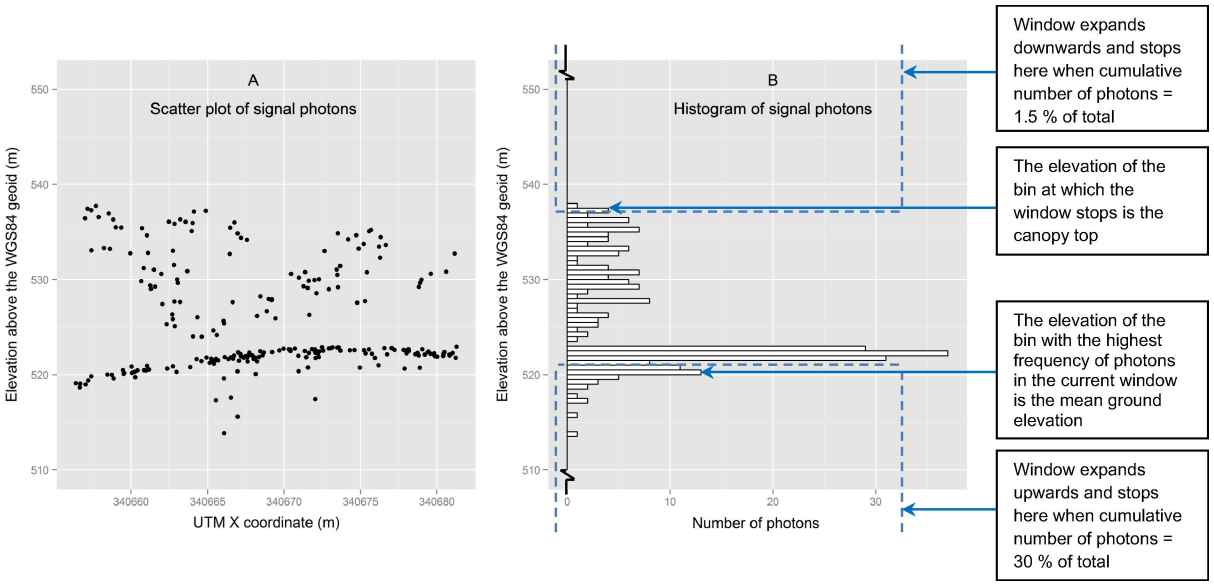


Figure 2.3: Diagrammatic representation of a tree height calculation algorithm [Gwenzi et al., 2016]

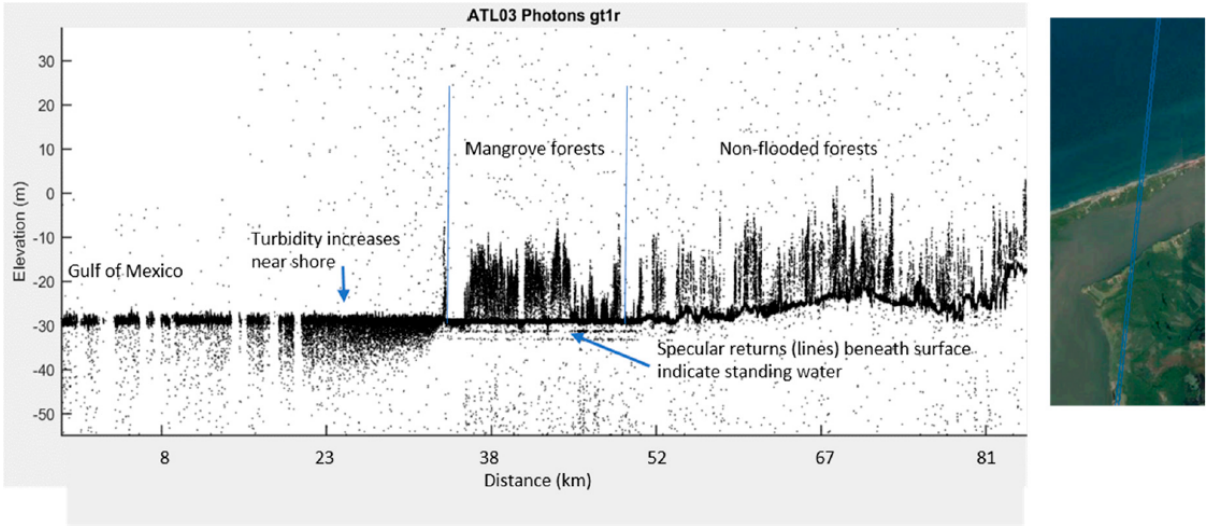


Figure 2.4: Profile of ICESat-2 over a mangrove forest in Mexico [Neuenschwander and Magruder, 2019]

Trying to understand the unique properties of each surface to be identified will undoubtedly be beneficial in the creation of the classification algorithm. From the research presented in this section, it is clear that the ATLAS signal behaves differently depending on the land type under illumination. Although there has been no published research specifically looking into multi-class land type classification of ATLAS data, there are various resources presented in this Section that have implied that there can be promising

results, not to mention the successful classification studies performed from the original ICESat mission. This poses a strong research direction to explore the use of ICESat-2 data for land type classification.

2.3 Random Forest Classification Algorithm

This Section provides a high level introduction to the random forest classification algorithm for the reader that is not familiar with its application. Section 2.3.1 provides an introduction to decision trees, from which random forests are constructed and Section 2.3.2 introduces the random forest classification algorithm and relevant attributes. For the readers familiar with the random forest classification algorithm, a discussion of the hyper-parameters chosen and test set-up can be found in Section 3.6.

2.3.1 Decision Trees

Decision trees are a predictive modelling tool widely used in machine learning. They are defined by recursively partitioning the input space, and defining a local model in each resulting region of input space. This is represented by a tree structure, where a set of questions (leaves) are hierarchically organised and where each "leaf" of the tree can be seen as its own sub-classification model.

Figure 2.5 provides a basic example of a decision tree. Figure 2.5a shows the decision tree structure and provides some terminology. Each node of the tree represents a question asked about specific properties of the input data. As the model is hierarchical, the question being asked depends on the answer to the previous question. All nodes have exactly one incoming edge and in the trees used in this report, two outgoing edges. The decision on the input sample is made in the terminal node/leaf, which is represented at the bottom of the tree and by the square leaves in Figure 2.5. This process is described in Figure 2.5b. The tree is attempting to determine properties about the image, in this case, whether the image has been taken inside or outside. The features or properties that are available for classification are the pixels in the image. One can start by asking whether the top part of the image is blue, if so, this indicates that a sky is present. Based on this, we can ask if the bottom part is also blue and if not, we have evidence that the scene is indeed an outside image. In simple problems, the parameters and structure of the trees can be selected by hand. However, for more complex problems or those where a large collection of features are used, the parameters and structure are automatically trained from the input data.

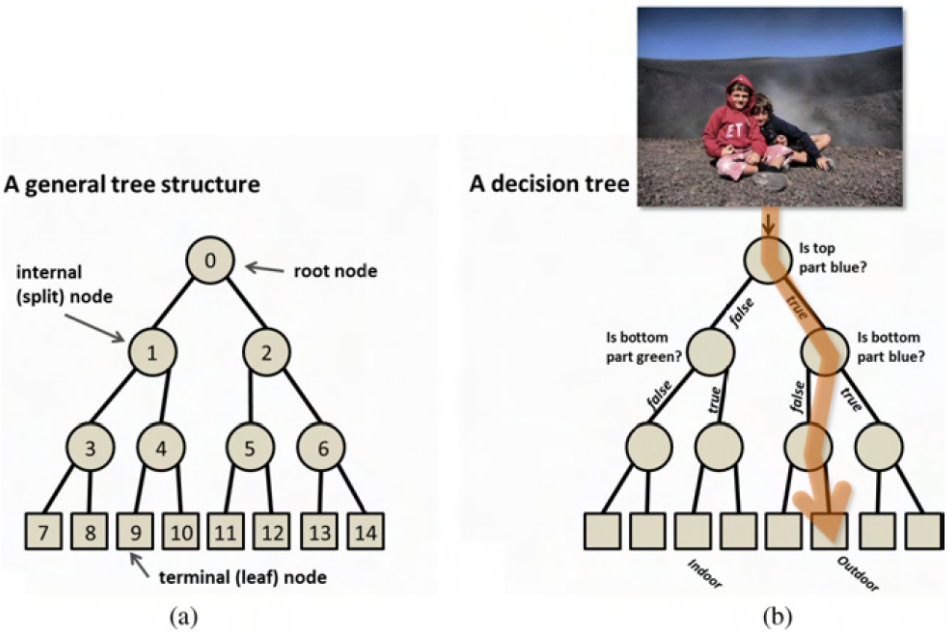


Figure 2.5: Decision Tree (**a**) a tree is a set of nodes or edges arranged in a hierarchical manner. (**b**) Each node (apart from a terminal node) stores a test/split function to be applied to the incoming data. Each leaf (terminal node) stores the final classification decision. [Criminisi et al., 2011]

The input sample to a decision tree is denoted by the vector $\mathbf{V} = (x_1, x_2, ..., x_d)$ where x_i represents a property of the input sample, commonly referred to in machine learning as a feature. Of course, the amount (dimensionality) and form of these features depend on the application at hand. Figure 2.6 shows a segment of a decision tree used in the research presented in this report. The features can be seen in the first row of each leaf. For example the root leaf contains the feature "half_perc", which indicates the ratio of photons that fall above and below a certain height. In each leaf, a test function (question) is defined, in this case, "Is the half_perc of the sample in question below or equal to 44.888? Depending on the answer, the sample moves through the tree to either of the samples below. In this example, it can be seen that of the $14,434$ samples entering at the root node, 7495 were True and 6939 False.

The input data is split into training and test data. As the training data is passed through the tree, a probability distribution is created at each test function (leaf). This is indicated in each leaf by the "value" attribute. The "value" attribute in each of the leaves indicates the amount of samples associated with each of the 5 classes used in the research, that is, if one were to terminate the tree at the depth indicated in Figure 2.6 (3), then the probability of a unseen sample belonging to a particular class would be calculated using the value attribute. For example, if an unseen sample travelled through the leaf and ended at the leaf on the bottom left, it would be classified as "Wetlands" as most (3757/8188) of the training samples landing in this leaf were Wetlands. As the true classes of the samples used in this report are known, it is considered a supervised classification task.

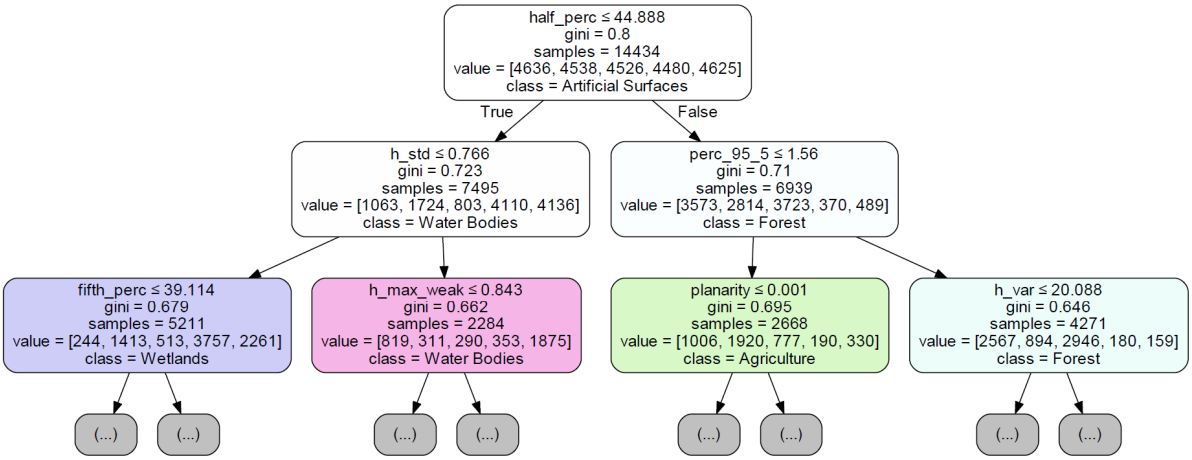


Figure 2.6: Decision Tree taken from classification applied in this research. For each leaf: Row 1 provides the test function, Row 2 the gini-importance, Row 3 the total amount of samples considered, Row 4 the distribution of these samples and Row 5 the predicted class according to that leaf.

Of course, the test functions are key to the successful application of the decision tree. The test functions are created during the training phase. At each node, $\mathbf{j}$, a subset, $\mathbf{S}_j$ of features of the input sample will be presented. From this subset, the test function that best splits $\mathbf{S}_j$ into $\mathbf{S}_j^{True}$ and $\mathbf{S}_j^{False}$ is learned, that is, there is a maximisation of an objective function at that node, according to Equation 2.1.

$$\boldsymbol{\theta}_j^* = \arg\max_{\boldsymbol{\theta}_j \in \mathcal{T}} I_j \tag{2.1}$$

where

$$\begin{aligned} I_j &= I\left(\mathcal{S}_j, \mathcal{S}_j^{\text{True}}, \mathcal{S}_j^{\text{False}}, \boldsymbol{\theta}_j\right) \\ \mathcal{S}_j^{\text{True}} &= \{(\mathbf{v}, \mathbf{y}) \in \mathcal{S}_j \mid h\left(\mathbf{v}, \boldsymbol{\theta}_j\right) = True\} \\ \mathcal{S}_j^{\text{False}} &= \{(\mathbf{v}, \mathbf{y}) \in \mathcal{S}_j \mid h\left(\mathbf{v}, \boldsymbol{\theta}_j\right) = False\} \end{aligned} \tag{2.2}$$

$\theta_j \in \mathcal{T}$ denote the parameters of the test function at the jth split and $\mathbf{y}$ is the known label/class of the input sample. I is the objective function, which measures the level of information gain at the split. In

order to determine the optimal split at the each node and according to Equation 2.1, a purity measure is utilised. This calculates the value at which the most discriminative information is gained from the split. This is described in Figure 2.7, where a simple example is presented using a purity measure called "gini-importance". Imagine you wanted to classify a group of 5 females and 5 males into their respective genders and that the average heights of females and males was $1.6m$ and $1.7m$ respectively. Assume that, each of the people is at or very close to average height. The gini-importance rewards values that better split the data and therefore is higher in the case where $1.65m$ was used as the discriminative value. The aim is to achieve a value at each node that maximised the gini-impurity at that node.

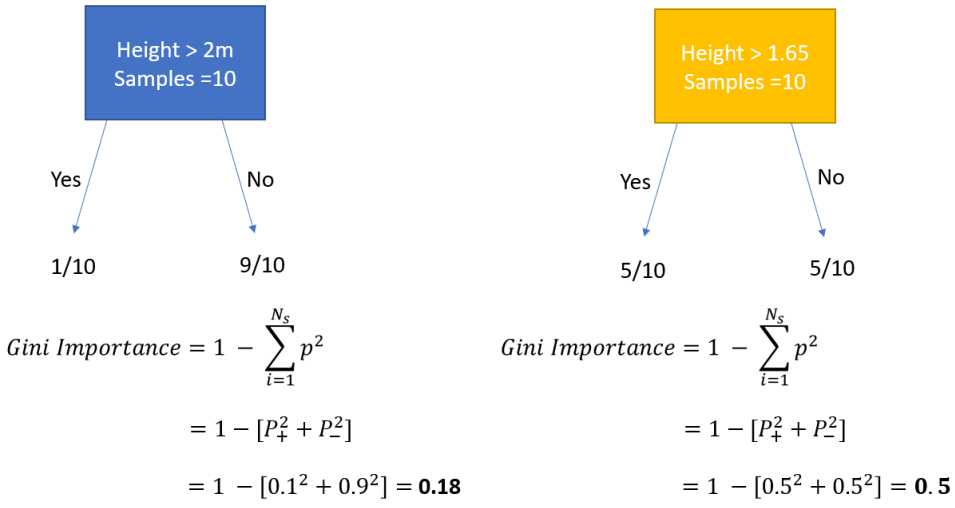


Figure 2.7: Example of gini importance. The goal is to maximise the importance at each split/test function. In this case, the value of $1.65m$ would be best because it splits the samples up and creates an information gain, that is, one gains more discriminative information about the sample.

The gini-importance, provided in Figure 2.3 and Equation 2.3 will be used as the objective function for the decision trees in this report.

$$Gini\ impurity = 1 - \sum_{i=1}^{N} p^2 \tag{2.3}$$

where N is the number of possible outcomes from the test function.

2.3.2 Random Forests

Random forests are a combination of decision tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [Breiman, 2001]. There are a few downsides to decision trees that are alleviated with the use of an ensemble of trees. Firstly, decision trees are sensitive to changes in the input data, meaning that a small change in the input data can have a significant effect on the optimal structure of the tree. Due to its hierarchical nature, the position of a feature in the tree will have a significant effect on the result below it. The key to the success of random forests is the fact that all of the trees are randomly different to one another. This leads to a de-correlation between the individual tree predictions, resulting in improved generalisation and robustness [Criminisi et al., 2011].

Classification Forests will be utilised in this research where the goal will be to match the data point, $\mathbf{V}$, to one of 5 discrete land types studied in this research. A discussion is now provided regarding some necessary attributes of the random forest algorithm. For all of the attributes discussed below, choices need to be made tailored to the application at hand. For this research, the choices and motivation are presented in Section 3.6.

Features

Features are a collection of properties extracted from the sample to be classified. For example, if one were classifying pictures of random animals, some discriminative features could be "Number of Legs", "Height", "Fur/No Fur" etc. One of the reasons as to why random forest is popular in classification tasks and the predominant reason as to why it was selected for this research is that the user chooses their own features (as opposed to, for example, neural networks, where the features are determined automatically). This is especially desirable in applications such as this, where one of the aims is to learn which attributes are best at discriminating between land types and which features are particularly distinct for each class. Random forest also deals well with a large amount of features. As a chosen random subset of features is chosen at each split, the computational time isn't effected by the the number of features. As a large number of trees are used, the effect of a poor feature on the overall performance is dampened as there will be many trees that will not consider that feature.

Number of Trees

This is simply the amount of trees in the forest. It has been shown that the testing accuracy of a forest increases monotonically with forest size [Criminisi et al., 2011]. Each tree in the forest will provide a probabilistic prediction as an output, that is, it will give the likelihood that a particular sample belongs to one of the five classes considered in this research. As the trees are all randomly different from one another, it follows that the generalisation of the forest would increase with an increasing forest size. In other words, the forest will have more of a capability of successfully classifying data it hasn't seen before. A single decision tree will tend to overfit and have a high variance/low generalisation, meaning that the model will be sensitive to small changes in the input data. These errors reduce as the amount of trees in the forest increases, however, the generalisation error does converge at some point as the forest becomes large [Breiman, 2001] and of course, the computational effort increases.

Tree Depth

This is the depth at which the trees of the forest are allowed to grow. For example, the decision tree in Figure 2.6 has a depth of 3. Of course, an increasing tree depth also increases the computation time. Intuitively, a tree that is too shallow will not have the resources to successfully discriminate between classes, leading to a poor overall classification accuracy. Conversely, a too large depth can lead to overfitting, as the tree has "learned" training data too well and will not be able to adapt to changes in the input data/ data it hasn't seen before. In general, the depth of the trees needs to be selected as a function of the model complexity.

Max Features

One of the powerful aspects of random forest is the injection of randomness into the model. One of the ways in which this is done is by selecting a subset, $\mathbf{S}_j$, from the total set of features to be considered for each tree. The maximum value of $\mathbf{S}_j$ controls the randomness of the forest. For example, if $\mathbf{S}_j = N_{features}$, then all trees have access to the same information and there will be no difference between them. The lower the value of $\mathbf{S}_j$, the more randomness introduced into the model. Again, the optimal choice is application dependant.

Bootstrapping and Bagging

Bootstrapping, also known as bagging, is another technique that introduces randomness into the model. Similarly, to "Max Features", a random subset of the available samples are used to build each tree. Once a tree has been built, the samples are replaced and another random selection is made for the next tree. This reduces the variance and helps reduce overfitting, improving the generalisation of the model.

To conclude, this section has provided the relevant background information to aspects dealt with in this report. Namely, an introduction to the ICESat and ICESat-2 mission in Sections 2.1 and 2.2 respectively. The promising classification results of ICESat-1 data combined with the level of detail that is apparent in ATLAS's vertical profiles provide evidence that they can be used to classify land types in The Netherlands.

The nature of the datasets provided by these missions are unique in the sense that they possess some of the attributes, negative and positive, of both "laser scanning data" and "satellite data". Furthermore, the GLAS and ATLAS datasets also possess certain attributes, positive and negative, that aren't typically attributed to these kinds of datasets. The ATL03 dataset is a unique product and kind of mix of a typical laser scanning and satellite product. For example, ATLAS is a laser scanning instrument, which would typically be expected to provide a dense cloud of photon locations. This is advantageous for applications such as the scanning of a bridge or similarly sized structure, or the creation of a Digital Elevation Model (DEM) over a relatively small area with use of an airbourne LiDAR. However, ATLAS does not provide the spatial density required to serve these purposes. However, while ATLAS provides low density return in comparison to a typical laser scanning product, it can provide this on the scale of the entire Earth, repeated every 92 days. Conversely, one would expect vast temporal and spatial coverage of the Earths' surface with a typical Earth Observation satellite, whereas ATLAS provides data only at the exact location of the laser footprints. These attributes place ICESat in a unique position to be able to study vertical profiles on a global scale.

The method that was chosen to attempt the classification was the Random Forest classification algorithm. An introduction to this subject is provided in Section 2.3 and there are two main reasons as to why it was chosen. Firstly, when using random forest one is able to select the features. This is of particular importance in this research, where one of the aims is to discover which features unique to the ATL03 dataset are most important. Secondly, most of the literature presented in this Section that was used for classification made use of either a decision tree or random forest. As the aim is not to test which classification technique is best, its seems a logical decision to continue this trend.

3. Data and Methodology

This Section will describe in detail, the methodology applied in this research as well as the data utilised to do so. The first 3 subsections aim to provide the reader with an introduction to the datasets used in the research. Section 3.1 provides a first look at the ATLAS data with a discussion of relevant characteristics. Section 3.2 presents the validation data and Section 3.3 presents the area over which the classification will be performed.

From Section 3.4, the reader is presented with the novel work of the thesis. Section 3.4 presents the research strategy, for example, What will actually be classified? How will the data be segmented? How will the results be validated? etc. Section 3.5 presents the features that were considered for the classification. For explanatory purposes these have been separated into three categories: "Altitude Derived", "Eigenvalue-based" and "Histogram-based", which are described in Sections 3.5.1, 3.5.2 and 3.5.3 respectively. Section 3.6 provides a description of the classification model and test set-up.

3.1 ATL03 Data

There are variety of data products provided by the National Snow & Ice Data Center (NSICD), a high-level description of which can be found in Figure 3.1. For the analysis performed in this project, the product "ATL03" will be used. As can be seen in Figure 3.1, ATL03 is one of the highest level data products and is the parent of most of the other data products.

ATL03, along with ATL04, are the base scientific data sets, having been derived, converted and formatted from the raw telemetry data. ATL03 comprises of latitude, longitude and elevation height for each recorded photon and ATL04 contains the atmospheric profiles of normalised relative backscatter.

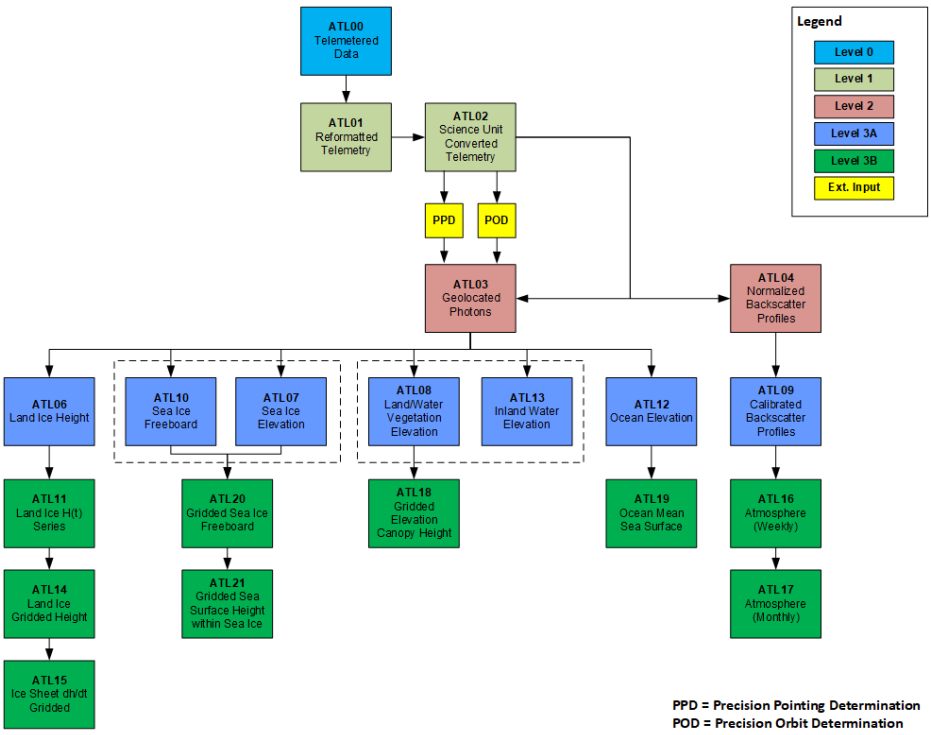


Figure 3.1: ICESat-2 data processing workflow and product map [NSIDC, 2020]

ATL03 will be used for the research presented in this report. One of the goals of the research is to determine to what extent the vertical profiles recorded by ATLAS can classify land types. For this reason, only the photon heights and their respective latitudes and longitudes are desired for the analysis, that is, the extra information calculated and provided in the lower level datasets in Figure 3.1 would be superfluous to the cause and move away from the goal of understanding the raw signal. As discussed in Section 2.2.1, ATLAS records individual photon returns, whether they have reflected off the surface under investigation, or from a cloud, or were just floating around in the atmosphere at the time of acquisition. As discussed in Section 2.2.1, for this research, only photons that have been classified with a "high" likelihood of being a signal photon have been used.

Figure 3.2 provides the ATL03 photon height profiles above an urban area, or Artificial Surface. The individual photons are represented by the points in the profiles and they have been connected in the Figure as it is easier to make out the underlying surface as well as patterns in the profile. The y-axis represents the height above the WGS-84 ellipsoid and the x-axis represents the along track path of the profile, where the laser footprint is represented by the red line passing across the image, below the profile. The profile directly follows the laser footprint, that is, if a vertical line were to be drawn from any of the photons to intersect with the laser footprint, that would be the location from the which the photon was reflected. The different colours in the profiles represent a different $100m$ segment from the associated image.

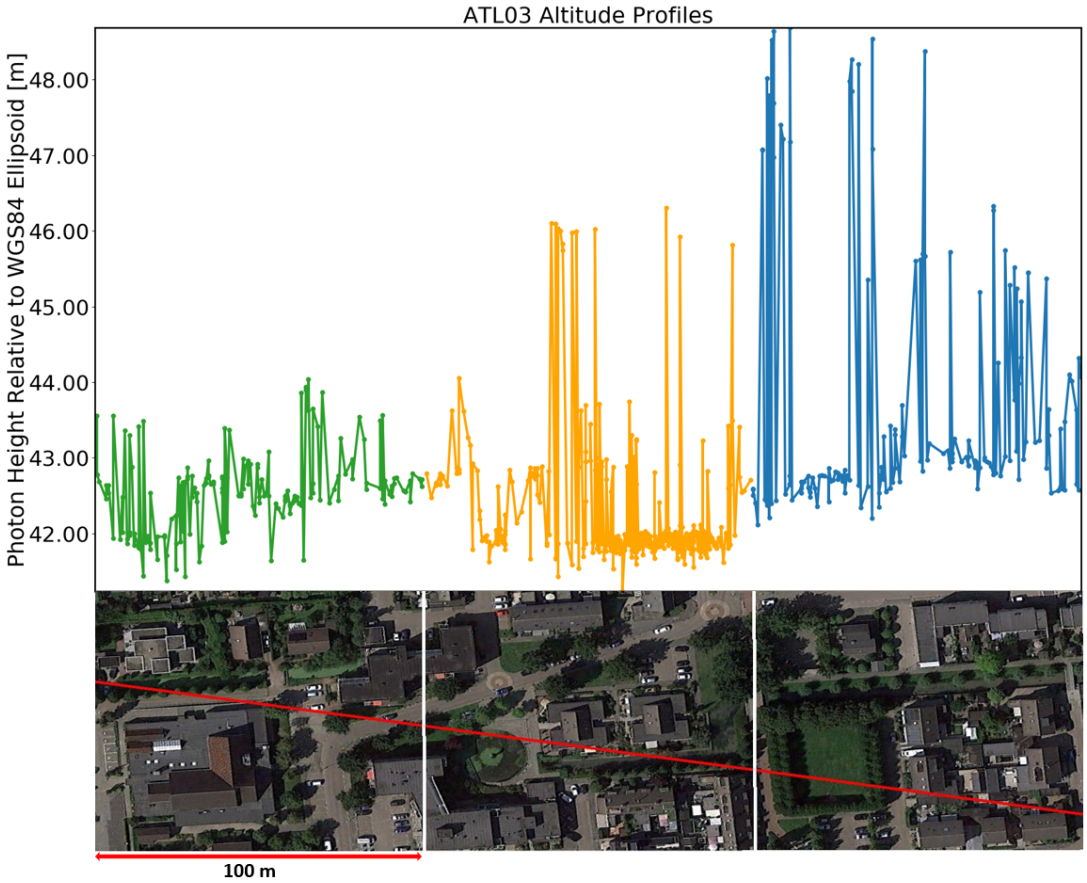
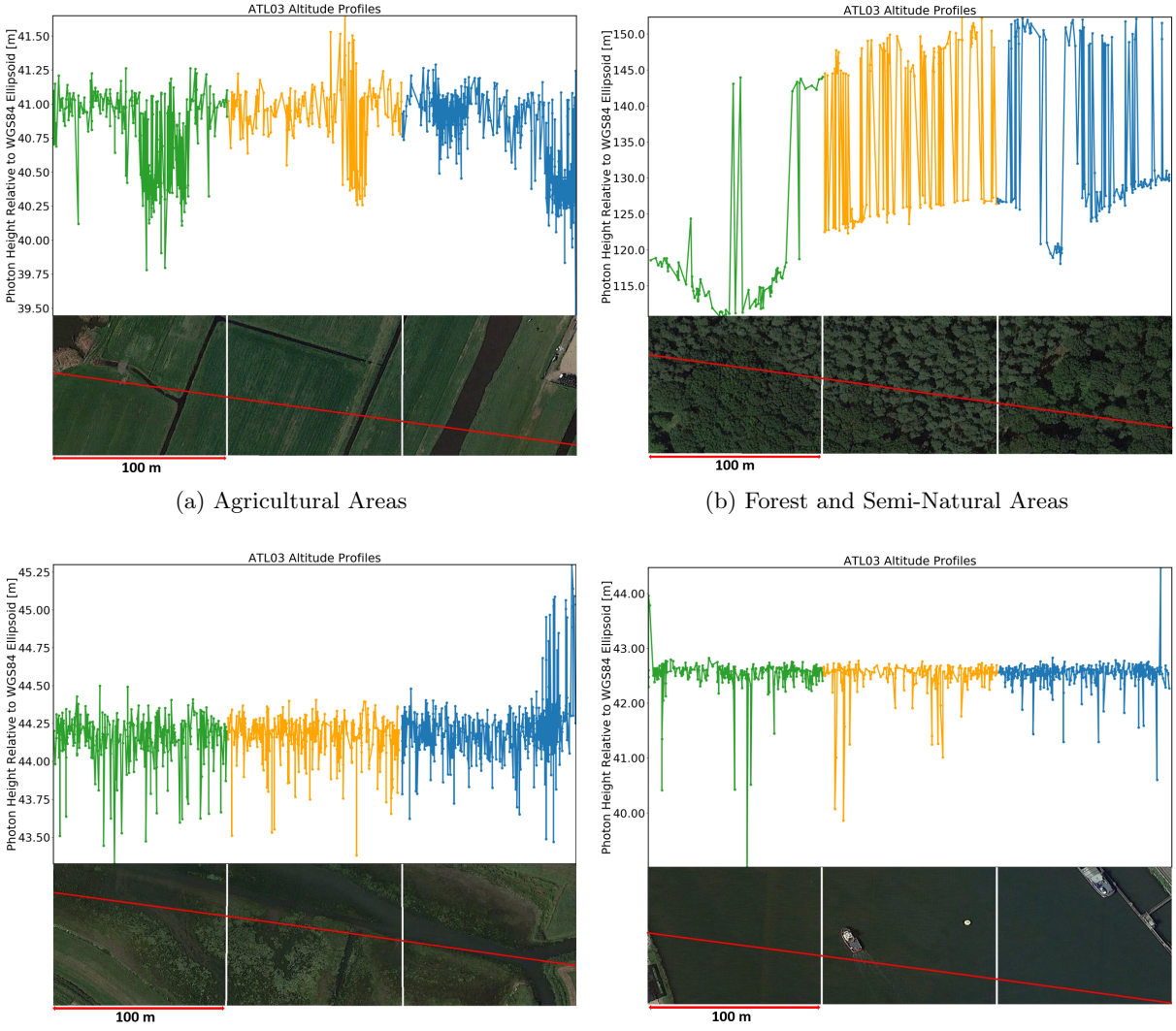


Figure 3.2: ATL03 Characteristic Photon Height Profiles above Artificial Surfaces.
The colours indicate a different $100m$ *segment as divided in the image. The laser footprint is shown by the red line passing over the Earths surface. If a vertical line were to be drawn from the photon in the profile to the laser footprint, the intersection would be the exact location that the photo was reflected from.*

It can be seen in Figure 3.2 that the Earth's surface is well represented in the photon returns. At the start of the third cell, there is a football pitch surrounded by trees. This can be seen traced out in the blue profile above it, where there are two dense areas of photon return, approximately $6m$ higher than the flat surface between them. Figure 3.3 provides characteristic profiles of the other 4 classes analysed in this report, that is, "Agricultural Areas", "Forest and Semi-Natural Areas", "Wetlands" and "Water Bodies".

Figure 3.3a shows the profiles over an Agricultural surface. The four small bodies of water that the laser passes over can clearly be seen by the dips in the profile. Agricultural areas tend to have a relatively bumpy profile, due to uneven ground or vegetation at varying heights. As a consequence, the underlying surface tends to be a little more difficult to make out than Artificial Surfaces or Forest and Semi-natural (Figures 3.2 and 3.3b).



(a) Agricultural Areas

(b) Forest and Semi-Natural Areas

(c) Wetlands

(d) Water Bodies

Figure 3.3: Characteristic Examples of ICESat-2 Vertical profiles above various land types. Due to the difference in range height of photon returns for different classes, the scale of the y-axis is not the same for the 4 profiles shown above.
The colours indicate a different 100m *segment as divided in the image. The laser footprint is shown by the red line passing over the Earths surface. If a vertical line were to be drawn from the photon in the profile to the laser footprint, the intersection would be the exact location that the photo was reflected from.*

In Figure 3.3b, the profile of a dense forest canopy is provided. Both the underlying surface and forest canopy are clearly defined, below and above trees of more than $20m$ in height. Note how the form of the profile when reflecting off a tree is the same as the trees around the football pitch in Figure 3.2. The form of the underlying surface in this example (Figure 3.3b) also implies a Forest and Semi-natural area class, as it is one that would unlikely be seen in the other classes, especially in the The Netherlands, which is mostly flat.

The final two characteristic profiles, provided in Figures 3.3c and 3.3d are Wetlands and Water Bodies. Water Bodies profiles possess the interesting characteristic that the surface of the water is sharply defined, accompanied with downward facing spikes, which are assumed to be reflected from within the body of water or ground surface below. Wetlands can largely be considered a mix of Water Bodies and Agriculture, as seen in the associated Wetlands image and its profile. It shares both the uneven return over a relatively small range of height, as with Agricultural surfaces, and the dense return and downward facing spikes found in Water Bodies.

It is clear to see the distinctive characteristics of the different land types in the examples provided in this Section. Of course, it is not always as clear, however these profiles signify the trends that are seen for each class throughout the data set. When discussing a "characteristic profile" in the remainder of this report, it will be referring to the attributes and characteristics that are easily detected visually from the examples provided in this Section.

3.2 Validation Data - CORINE Land Cover Database 2018

The CoOrdination of INformation on the Environment (CORINE) Land Cover (CLC) database is a European wide land cover database that has been maintained and updated since 1986. The database consists of an inventory of land cover in 44 classes, separated into 3 levels of detail. The highest level classes are "Artificial Surfaces", "Agricultural Areas", "Forest and Semi-Natural Areas", "Wetlands" and "Water Bodies". These are also the land types classified in this research.

The most recent version, "CLC2018", was produced from Land cover observed in 2018, predominantly via Sentinel-2 data. The land cover is provided at 100m resolution. Figure 3.4 provides the CLC2018 over The Netherlands. CLC2018 can be accessed at [Copernicus Land Monitoring Service, 2019a]. The minimum width of the linear mapping elements in the dataset is $100m$, that is, each pixel presented in Figure 3.4 is 100x100m grid cell. Each cell is associated with 1 of 44 land types.

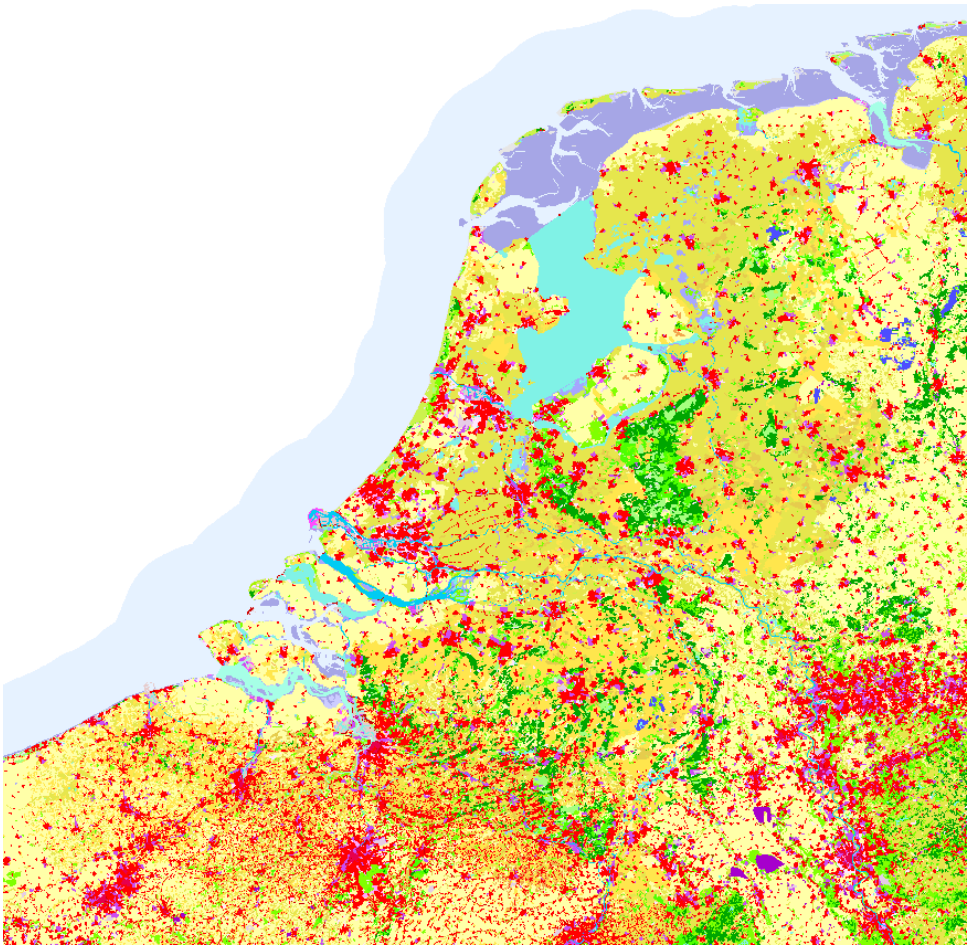


Figure 3.4: CORINE 2018, The Netherlands. Each pixel is $100x100m$ and each colour represents a different land class

Figure 3.5 represents the land cover class breakdown. It was decided to use the highest level land cover types according to the CLC2018 for this research. This is due to the fact that multi land type classification of The Netherlands with the use of ATLAS data has not been performed before and therefore the capability is unknown. [Copernicus Land Monitoring Service, 2019b] provides a detailed discussion of what constitutes the classes presented in Figure 3.5.

Its necessary to be aware of the scope of land types within the classes, that is, some classes constitute a much wider variety of surface types. For example, Water Bodies basically consist of rivers, estuaries

and the ocean/sea. One would not expect particularly large variations of profiles between these sub-classes. On the other hand, Forest and Semi-natural areas consists of all forest types, beaches and surrounding areas, as well as natural grassland and sparsely vegetated areas. The surface variations from certain classes, such as Forest and Semi-natural areas, is larger and will likely present difficulties in the performance when compared to a class such as Water Bodies which is anticipated to have a low surface variation across its profiles.

Figure 3.5: CLC2018 Land Cover Class Breakdown.
Only the highest level of classes will be used in this research, that is, "1. Artificial Surfaces", "2. Agricultural Areas", "3. Forest and Semi-Natural Areas", "4. Wetlands", "5. Water Bodies"[Copernicus Land Monitoring Service, 2019b]

3.3 Study Area

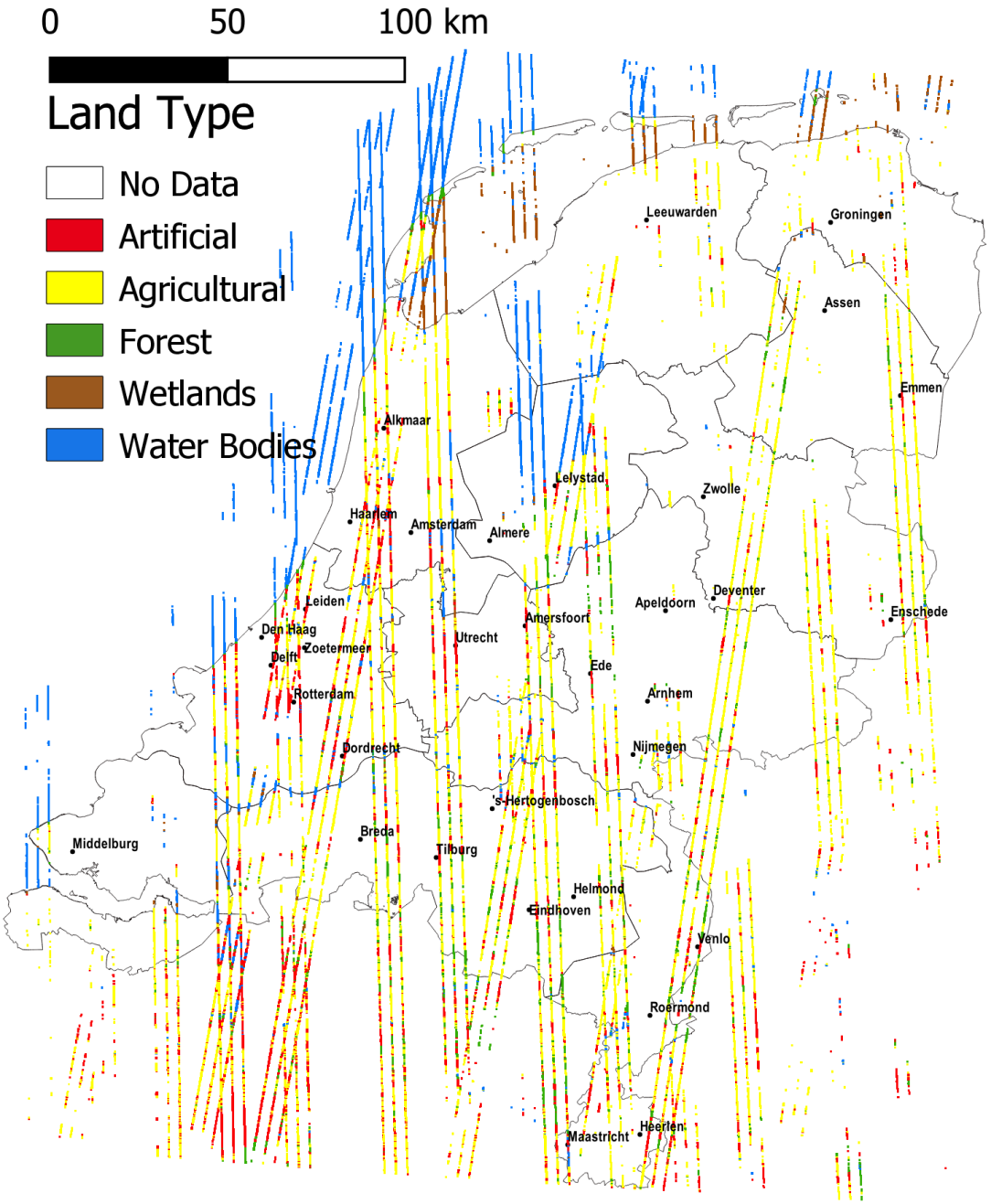


Figure 3.6: ICESat-2 Laser Footprints and associated land tyes - The Netherlands
Acquired between September 26th - December 26th 2019

The study area used in this research is a grid in Western Europe that encapsulates The Netherlands. While The Netherlands was the focus of this research, some data is also acquired above Belgium and Germany. Of the 1387 distinct ground tracks per 92 day repeat orbit of ICESat-2, 30 pass over The Netherlands and the data collected from these 30 ground tracks are used in the analysis. The data is taken from ICESat-2's fifth repeat orbit (Cycle 5) and is taken between September 26^{th} 2019 and December 26^{th} 2019.

The exact study area can be seen in Figure 3.6, where all the data used in this research is shown above the study area. The colours indicated the land type at that location. The tracks in Figure 3.6 actually represent the exact locations in which data was considered. The along track gaps in the footprints indicate that it was considered too cloudy to include the data at that location and time. The coloured pixels indicate the CORINE land type at that location.

3.4 Land Type Classification Strategy

The ICEsat and ICESat-2 datasets were introduced in Sections 2.1 and 2.2 respectively. Both ICESat missions share similar properties, however there are some significant differences in the data collection that instigates a different approach to the classification. The original ICESat mission shot laser pulses, which resulted in elliptical footprints where the distance between the footprints (where no data was acquired) being proportional to the pulse frequency of GLAS. It will not work to follow the same approach as [Duong, 2010], who classified each of these elliptical laser pulse footprints, as the ICESat-2 photon returns are essentially continuous. In this sense, it is necessary to partition the continuous photon return provided by ATLAS and create a custom classification/validation strategy.

An example of the ATLAS footprints can be seen in Figure 3.7, where photon signal data is acquired at only and exactly the locations of the laser footprints. Figure 3.7 (left) shows the footprints of one pass over The Netherlands, where the different colours simply indicate 1 of 6 laser tracks. Figure 3.7 (right) shows the zoomed in view of the footprints over Amsterdam, The Netherlands. It was decided to start by classifying square grid cells in accordance with the CLC2018 coordinate system. A description of and motivation for choosing the CORINE dataset can be found in Section 3.2. A grid was modelled over The Netherlands in accordance with CLC2018, that is, the 'epsg:3035' coordinate system [epsg.io, 2019].

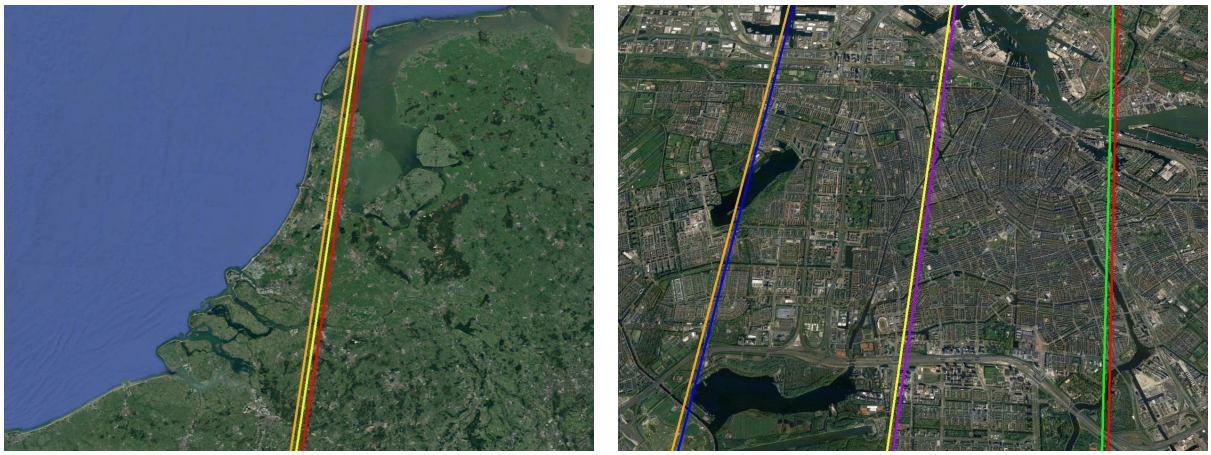


Figure 3.7: Examples of ATLAS Laser tracks over (left): The Netherlands and (Right): Amsterdam
Different colours indicate a different laser track (6 per track). In total there are 30 Reference ground tracks passing over The Netherlands, that is 30*x*6 *unique laser footprints.*

Figure 3.8 provides the modelled classification grid and ATLAS footprints superimposed over Amsterdam. Of course, one is only able to classify the land at the locations where the footprints interact with the surface, therefore the grid cells to be classified are only those in which the laser footprints pass through. This is depicted in Figure 3.9, where only the cells to be classified are shown. As can be seen, the vast majority of grid cells do not coincide with the ATLAS footprints and will therefore not be used in the classification. Every individually recorded photon has been provided with the respective coordinates in space at which it was reflected, and by converting these to the coordinate system of CLC2018 (epsg:3035), each photon was assigned to a grid cell, with the collection of all photons lying in a cell to be used for the classification.

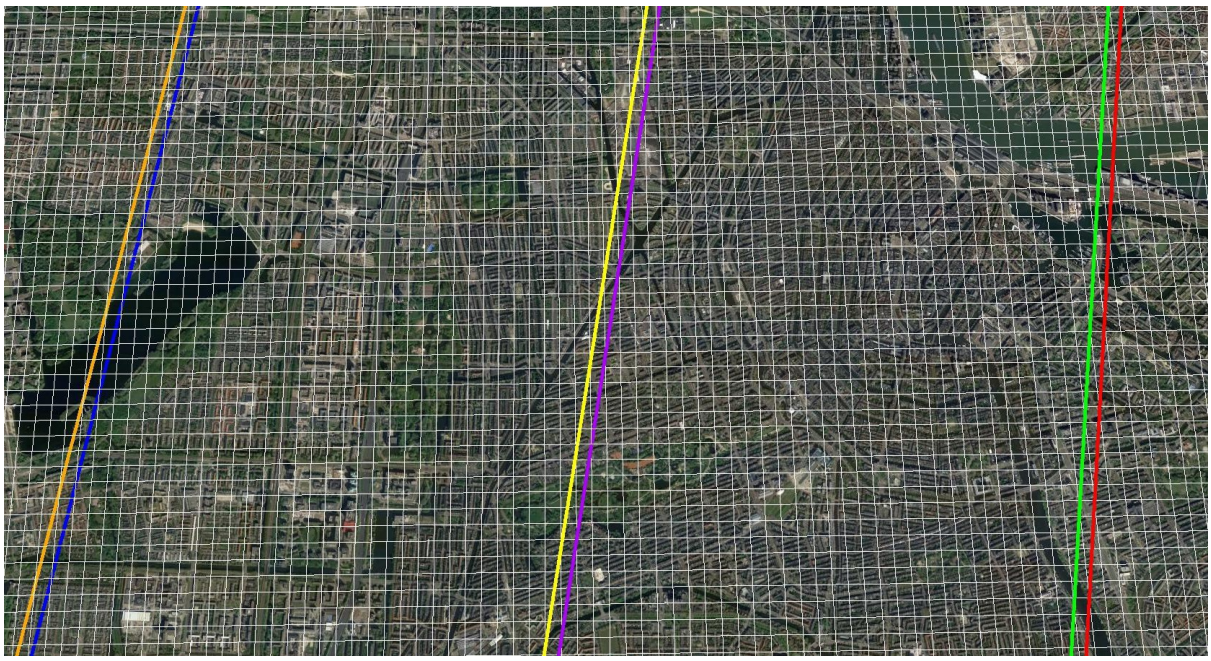
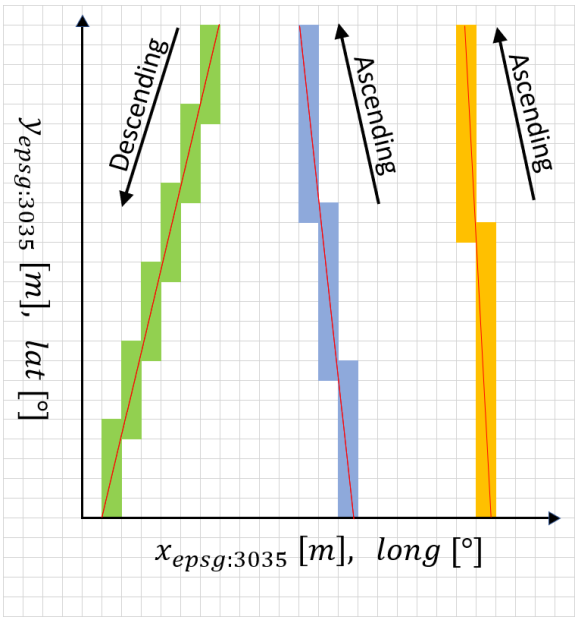


Figure 3.8: Classification Grid, Amsterdam
Each cell is $100x100m$*, aligned with the same coordinate system as the CORINE validation data, that is, the "epsg:3035" coordinate system.*

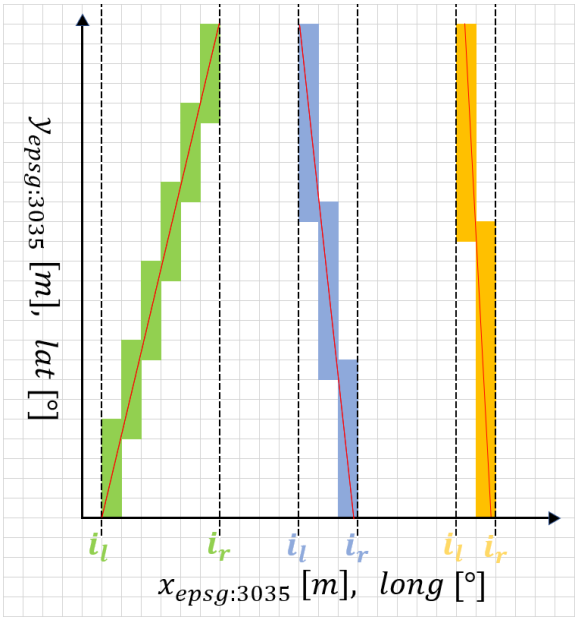
Figure 3.9: Reduced Classification Grid, Amsterdam, The Netherlands.
The areas where the laser tracks can be seen without a classification cell did not meet the minimum number of photons (50) to be considered for a classification cell

Due to the number of CORINE grid cells lying over The Netherlands (13,246,236) as well as the amount of photons being returned by each laser track ($>> 10^6$), it was important to remain computationally efficient when assigning photons to grid cells. Figure 3.10 provides a graphical description of how this was executed by processing as few grid cells as possible. As can be seen in Figure 3.9, the cells in which the tracks pass through form a recurring pattern across the Earths' surface. Following the laser tracks

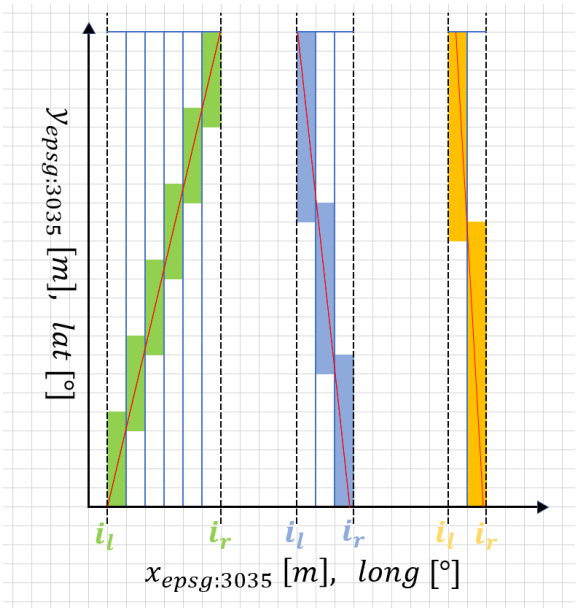
from the bottom to the top of the image (South to North), it can be seen that for every one cell in the Easterly direction, the pattern moves 6 or 7 cells in the Northerly direction. This pattern holds for all of the laser tracks considered in this research, that is, for every 1 cell in the east-west direction, the pattern moves > 1 cell in the north-south direction. This is further exemplified in Figure 3.10a, where 3 different laser track patterns are presented. Depending on the reference ground track, the orbit of the satellite is either 'ascending' or 'descending', that is, travelling in the south-north or north-south direction.



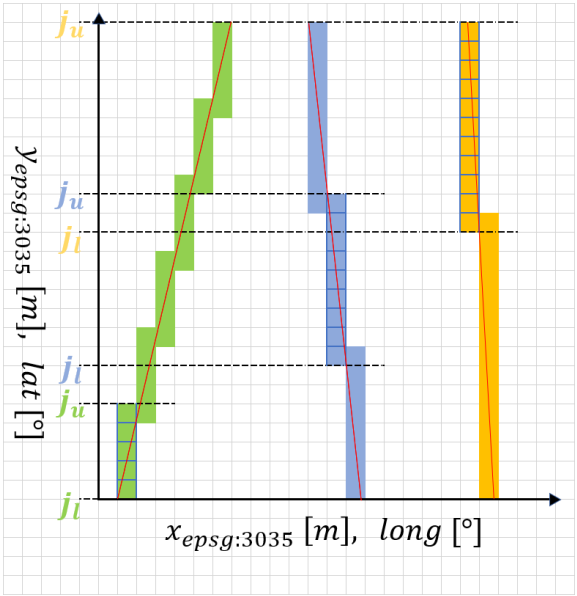
(a) Ground Track Grid Cell Pattern Variability



(b) Obtaining x/longitude bounds of laser tracks



(c) Grid/Column creation in x/longitudinal direction



(d) Grid/Row creation in y/latitudinal direction

Figure 3.10: Graphical Description of Photon Assignment Algorithm

The first step, depicted in Figure 3.10b is to obtain the bounds in the x (longitudinal) direction. This is achieved by searching the photon locations for the extreme values and then creating a grid from these values rounded to the nearest $100m$, in accordance with the 'epsg:3035' coordinate system. The start

and end point of the grid in the x-direction are labelled i_l and i_r respectively. The grid now consists of $(i_r - i_l)$/grid_resolution columns, as shown in Figure 3.10c. The algorithm loops through each of these columns and extracts all photons that lie within the bounds of the column. The procedure is now repeated in the j (latitudinal) direction. First, the bounds are obtained, again by rounding up to the nearest $100m$ for the upper bound and down for lower, creating a grid between the minimum ($j_l = y_{min}, j_u = y_{max}$) and maximum useful grid cells for that column. This is depicted for each laser pattern in Figure 3.10d, where the highlighted grid cells are the only cells that will be searched for that particular column. One can see that by applying the algorithm in this order, the minimum amount of grid cells are considered.

Once all the photons in a track have been assigned to a cell, an object dictionary can be made of all the relevant cells, where all the necessary information pertaining to that cell can be stored, such as the name of the cell, the photon data, the name of the reference ground track etc. All the information needed to create the features used for classification will be stored in the dictionary. The features that are made from this information will also be stored in this dictionary.

3.5 Data Attributes and Feature Extraction

This Section will present the features that were considered for the classification. In this report, a feature is defined as a measurable property or characteristic of a group of photons bound within a classification cell. For explanatory purposes, the features have been separated into 3 classes, namely, "Altitude Derived", "Histogram-based" and "Eigenvalue-based". A description of each class as well as its associated features can be found in Sections 3.5.1, 3.5.2 and 3.5.3 respectively.

The features were created based on a mixture of the use of previous research and novel solutions. The "Altitude Derived" features consist of mainly of simple statistical measures of the photon heights, such as averages, standard deviation and percentile heights. It has been shown in that simple statistical measures such as these could provide discriminative information regarding certain land types [Liu et al., 2020][Neuenschwander and Magruder, 2019][Ni et al., 2017].

The histogram-based features, derived from scattering surfaces discovered from histograms of the photon heights, are novel to this work. They were created in an attempt to define multiple scattering surfaces within classification cells. It was hoped that this would improve the error rate between classes such as Forest and semi-natural areas and Artificial. Upon first look, the profiles of these two classes look relatively similar, however, the strength, location and relative location of their scattering surfaces are distinctive.

While the laser footprints seem to directly trace the ground track of the satellite, it is stated that the footprint has an across track variation of $\approx 14m$ [NSIDC, 2019]. As this is is the same order of magnitude as the classification cells, it seems logical that there could be discriminative information in this dimension. Moreover, this would become more significant as the size of the classification cell became smaller. It was therefore decided to implement features that considered all three dimensions and are henceforth referred to as Eigenvalue-based features. There is vast literature covering the extraction of geometric features from point cloud/laser scanning data. The analysis presented here mainly focused on the work presented in [Weinmann et al., 2015][Demantké et al., 2011].

Extensive feature testing wasn't performed in the creation of the classification model, as one may expect during the training phase. Firstly, as discussed in Section 2.3.2, feature optimisation is not necessary as the overall effect of a poor feature is spread across the many trees in the classification forest. As the aim of this research is not to optimise the classification result but observe which features are significant or not, it was deemed sufficient to study the feature importance in the testing phase of the algorithm. This is discussed in further detail in the feature importance results, Section 4.2. There were more ideas for features that are discussed in the recommendations, Section 5.2. These were not implemented due to complexity, time constraints or falling out of the scope of this report.

All of the features are calculated directly from the elevation profiles of the classification cells, that is, from the altitudes of the individual photons recorded within the cells. The features presented can be created using any individual or combination of photon signal classifications (i.e 'low', 'medium', 'high', 'noise').

It is necessary to mention that the features presented in this Section are by no means the full set of potential features that were considered or that were thought capable of contributing toward the discrimination of the land types. However, in keeping with the global nature of the dataset, it was desired to create functionality that could classify large areas of land, such as The Netherlands. In order to fulfil this, it is required to remain computationally efficient in the creation of features. Therefore, the features that have been chosen are all calculable in a few mathematical operations. As an example, it was thought that the co-efficients of varying degrees of polynomials used to fit the altitude profiles would have the potential to characterise the land classes. However, this was also considered unfeasible with respect to processing time when considering the $> 10^6$ classification cells above The Netherlands and keeping in line with the desired computational efficiency.

3.5.1 Altitude Derived Features

The altitude derived features consists of 22 simple statistical properties taken directly from the raw photon heights. The calculations were performed over the groups of photons assigned to the classification cells. The features are depicted in Figures 3.11 to 3.16. For comparison, Figures 3.11 to 3.14 show arbitrary profiles of the different classes, with the altitude derived features depicted.

Mean Height

The average height above the WGS84 ellipsoid of all the photons within the classification cell. Depicted in Figures 3.11 and 3.12 by the solid line, labelled h_{mean} and calculated using Formula 3.1:

$$h_{mean} = \frac{\sum_{i=1}^{N_{ph}} h_i}{N_{ph}} \tag{3.1}$$

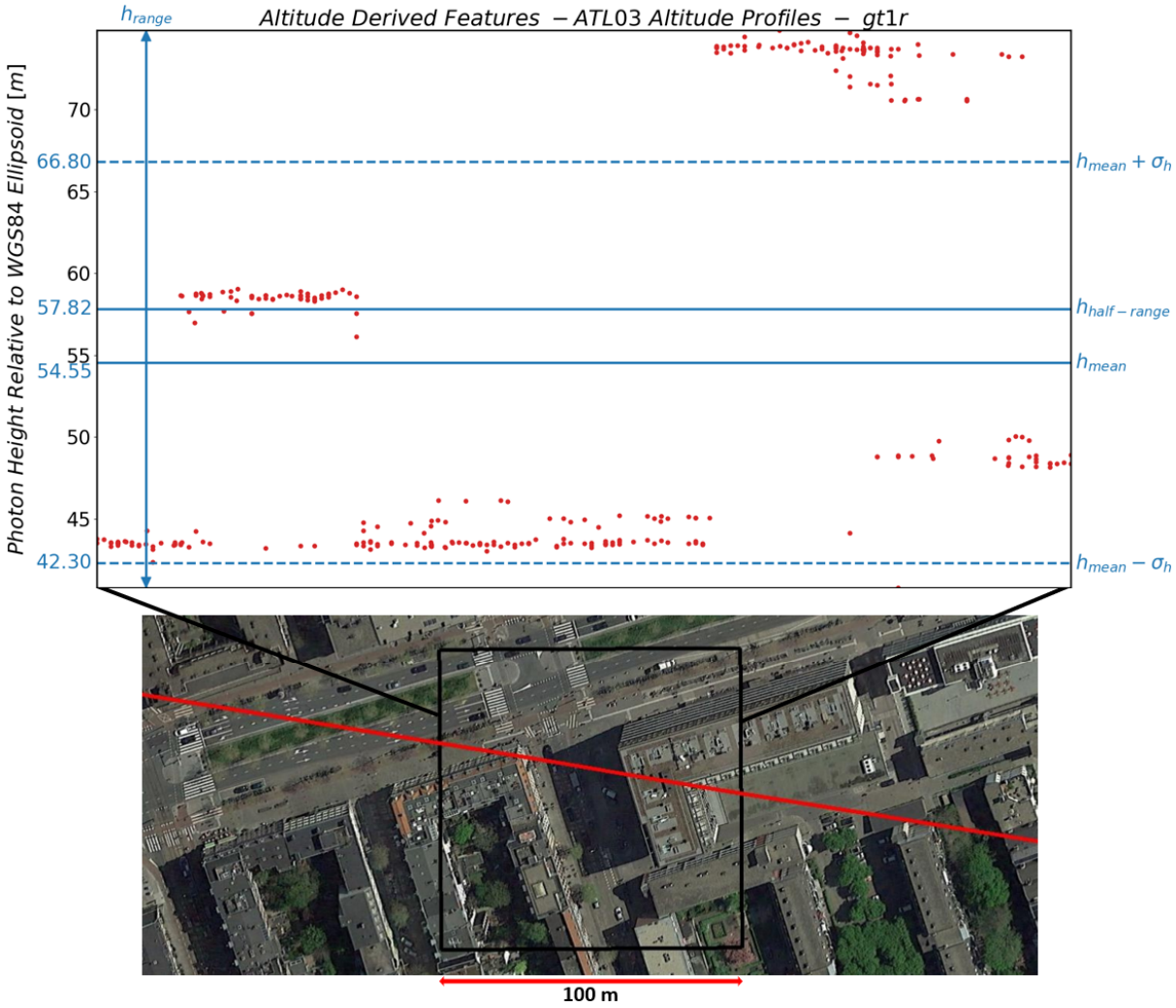


Figure 3.11: Altitude Derived Features for the classification cell bounded within the black border - Artificial Surfaces

h_{range} is equal to the range of the $y-axis$. h_{mean} and $h_{half-range}$ are shown with the solid blue line and the standard deviation with the dashed blue line.

Standard Deviation
The standard deviation of the photon heights above the WGS84 ellipsoid of all the photons within the classification cell. Depicted in Figures 3.11 and 3.12 by the dashed line, labelled $h_{mean} + \sigma_h$ and $h_{mean} - \sigma_h$ and calculated using Equation 3.2:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N_{ph}} (h_i - h_{mean})^2}{N_{ph} - 1}} \tag{3.2}$$

Variance
The variance of the photon heights above the WGS84 ellipsoid of all the photons within the classification cell, calculated using Equation 3.3

$$var = \frac{\sum_{i=1}^{N_{ph}} (h_i - h_{mean})^2}{N_{ph} - 1} = \sigma^2 \tag{3.3}$$

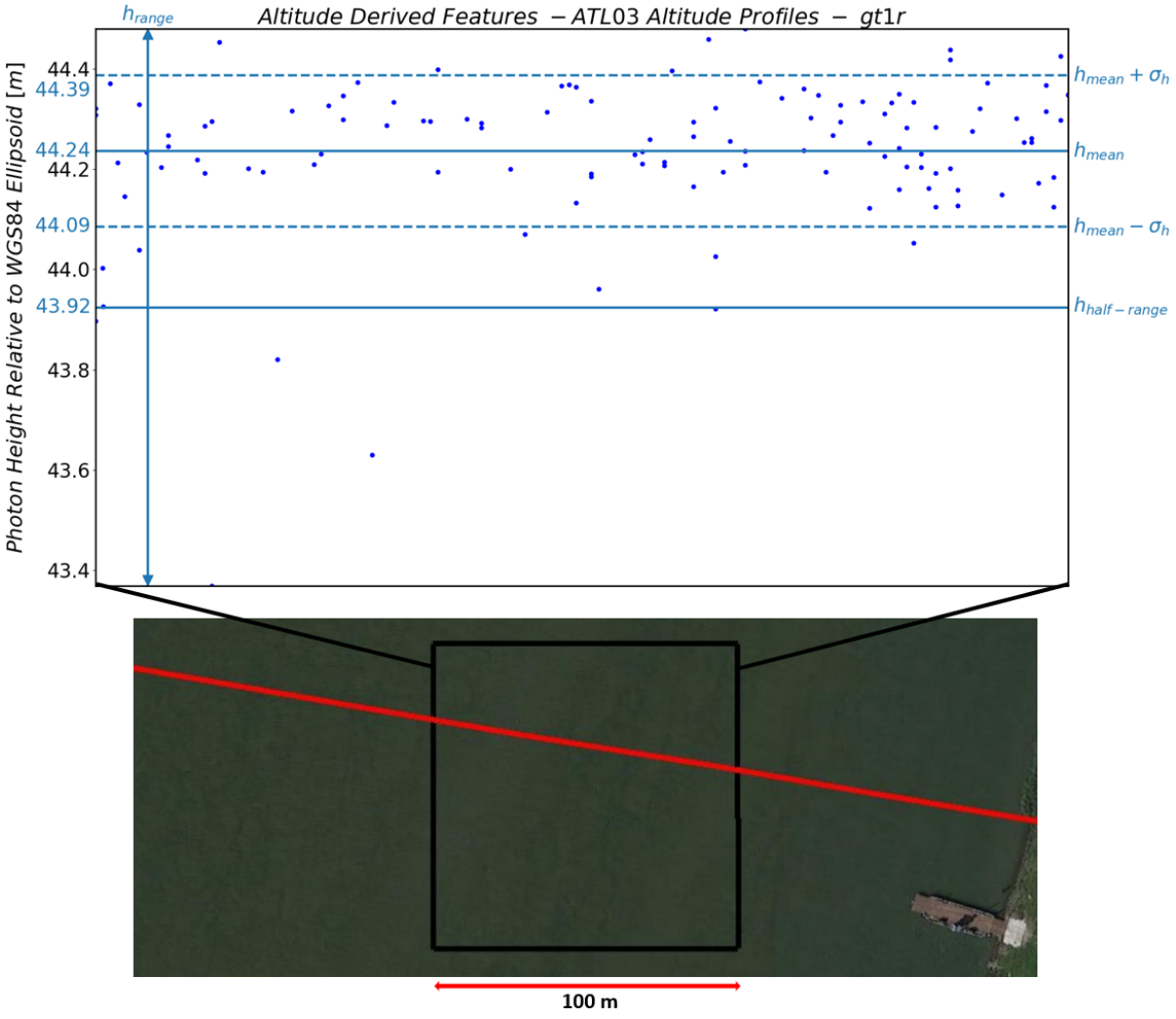


Figure 3.12: Altitude Derived Features for the classification cell bounded within the black border - Water Bodies
h_{range} is equal to the range of the $y-axis$. h_{mean} and $h_{half-range}$ are shown with the solid blue line and the standard deviation with the dashed blue line.

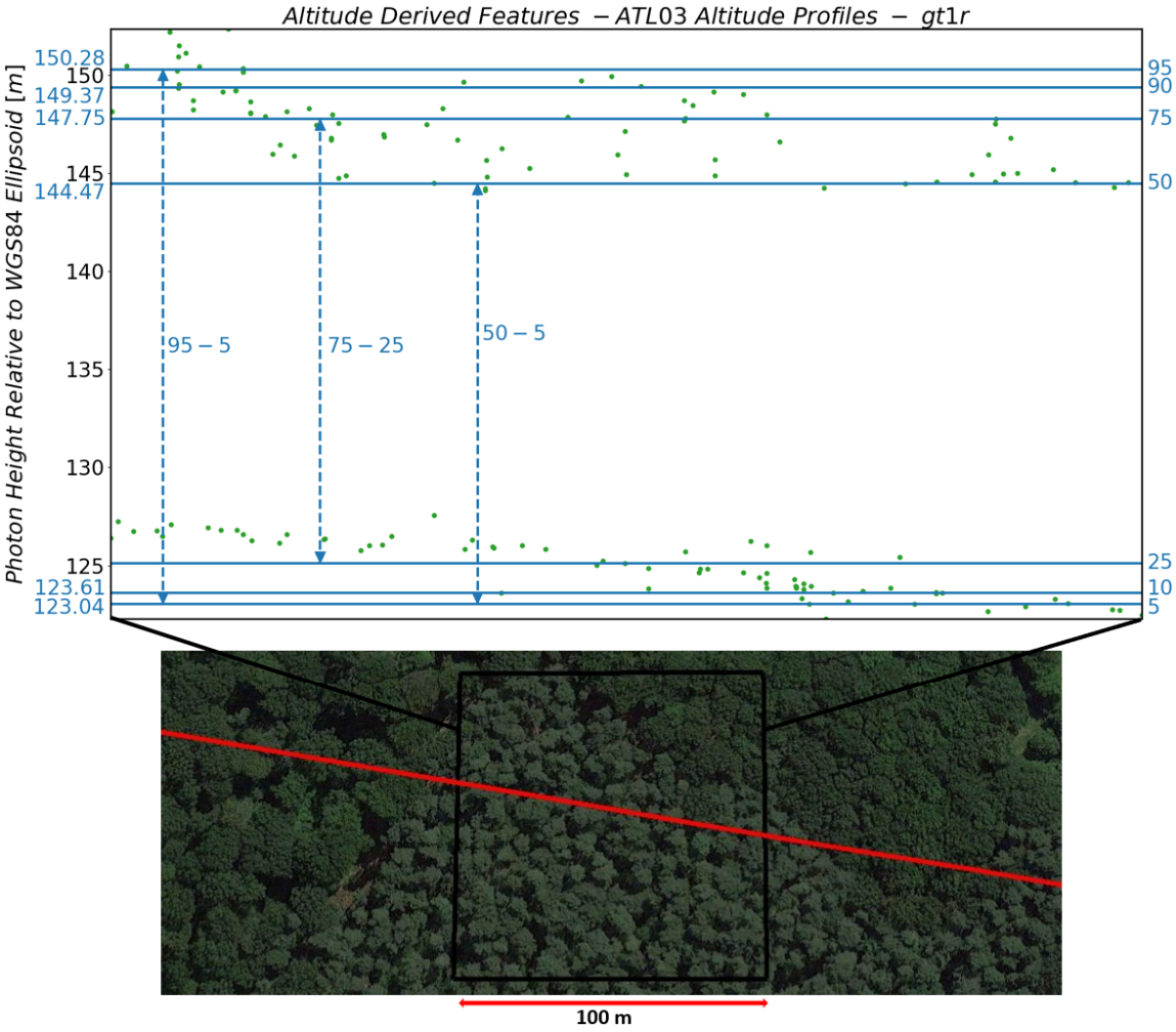


Figure 3.13: Altitude Derived Features for the classification cell bounded within the black border - Forest and Semi-natural areas

The 5, 10, 25, 50, 75, 90 and 95 height percentiles are represented by the solid horizontal lines. The 95-5, 75-25 and 50-5 percentile ranges are represented by the dashed vertical blue lines.

Minimum Height

The lowest height of all photons within the classification cell. Depicted in Figures 3.11 to 3.14 as the lowest point on the y-axis and calculated using Equation 3.4.

$$h_{min} = min(h_{ph}) \tag{3.4}$$

Maximum Height

The maximum height of all photons within the classification cell. Depicted in Figures 3.11 to 3.14 as the highest point on the y-axis and calculated using Equation 3.5.

$$h_{max} = max(h_{ph}) \tag{3.5}$$

Range Height

The height range of the photons above the WGS84 ellipsoid of all the photons within the classification cell. Depicted in Figures 3.11 and 3.12 by the solid vertical blue line, labelled h_{range} and calculated using Equation 3.6.

$$h_{range} = h_{max} - h_{min} \tag{3.6}$$

Half Range Ratio

The Half Range Ratio (HRR) is the ratio of photons above and below the mid point between the lowest and highest photon. Depicted in Figures 3.11 and 3.12 as the number of photons above and below the $h_{half-range}$ line and calculated using Equation 3.7.

$$HRR = \frac{\sum_{i=1}^{N_{ph}} h_i}{\sum_{i=1}^{N_{ph}} h_j} \tag{3.7}$$

where

$$h_i > h_{min} + \frac{h_{range}}{2},\ h_j \leq h_{min} + \frac{h_{range}}{2} \tag{3.8}$$

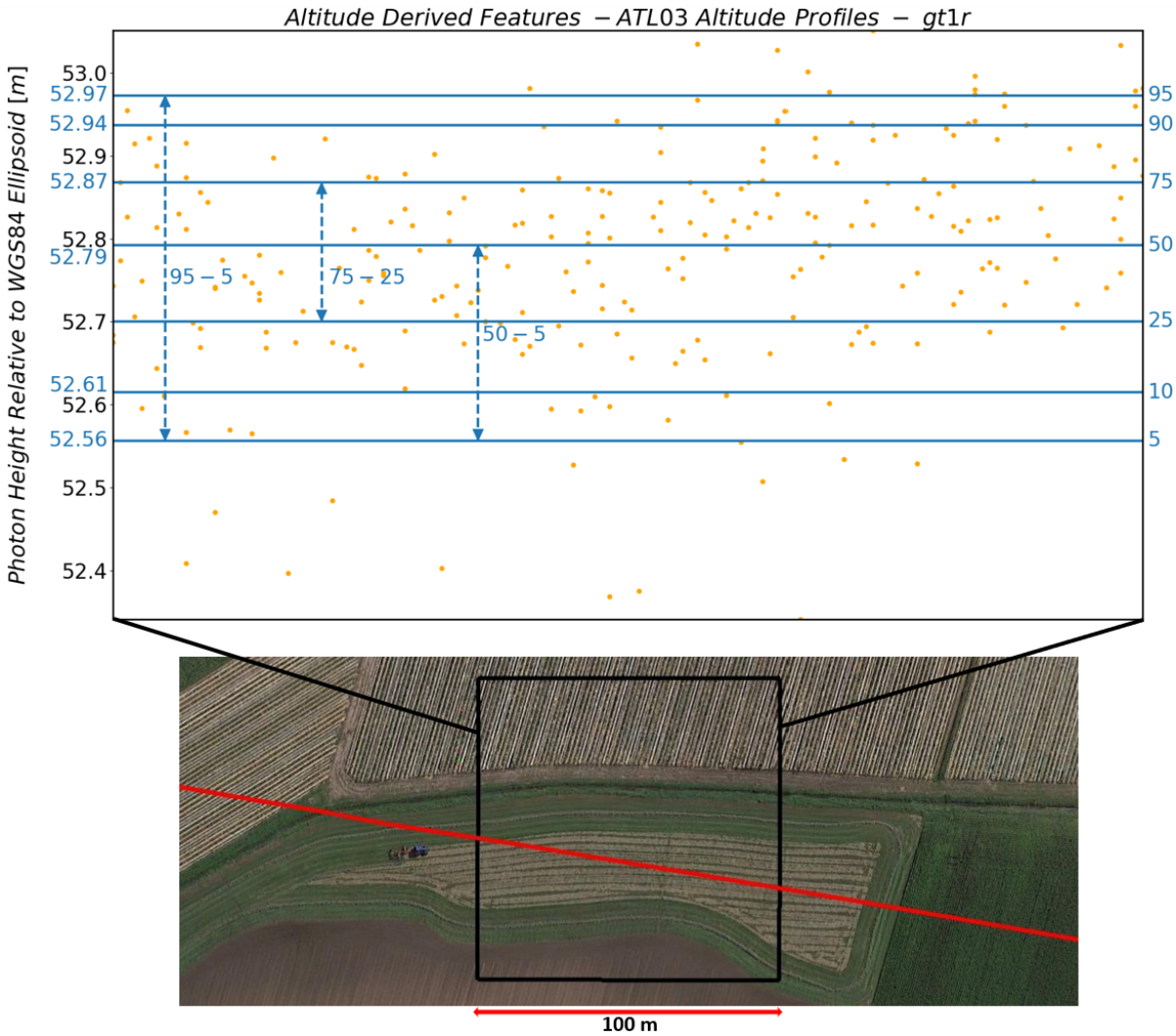


Figure 3.14: Altitude Derived Features for the classification cell bounded within the black border - Agricultural Areas

The 5,10, 25, 50, 75, 90 and 95 height percentiles are represented by the solid horizontal lines. The 95-5, 75-25 and 50-5 percentile ranges are represented by the dashed vertical blue lines.

Normalised Height
The normalised height is the lowest height subtracted from the mean height. Calculated using Equation 3.9.

$$h_{normal} = h_{mean} - h_{min} \tag{3.9}$$

Percentile Heights
A variety of percentile heights were used as features, that is, the height at which $x\%$ of the photons fall below. The percentile heights were calculated at 5%, 10% ,25%, 50%, 75%, 90% and 95%, depicted in 3.13 and 3.14 by the solid lines labelled with the respective percentile. Percentiles are calculated by ordering the photons heights and taking the height at which n photons fall under. n is calculated from Equation 3.10.

$$n = (\frac{P}{100}) * N_{ph} \tag{3.10}$$

where P is the value of the desired percentile.

Percentile Ranges
The range difference of percentile heights, depicted in Figures 3.13 and 3.14 by the vertical dashed lines for the $95-5$, $75-25$ and $50-5$ percentile ranges. The percentile ranges are calculated using Equation 3.11.

$$h_{p-range} = h_{p_1} - h_{p_2} \tag{3.11}$$

where $P_{1,2}$ represent the values of the desired percentile heights to find the range between.

Photon Reflectance/ Apparent Surface Reflection (ASR)
The ASR is a measure of the received laser pulse energy divided by the transmitted laser energy multiplied by the two way atmospheric transmission. The ASR is always modified by the atmospheric transmission, which is in general unknown for a given location and time, or at least difficult to obtain. ASR is calculated using Equation 3.12:

$$\rho = \frac{\pi N_{ph} r^2 D_c F}{N E A_t S_{ret}} \tag{3.12}$$

where N_{ph} is the number of photons received, r is the distance between the satellite and the surface, D_c is the detector dead time correction factor, F is a calibration factor, E is the outgoing laser pulse energy, A_t is the area of the telescope, S_{ret} is the product of the transmittance of the optics and the quantum efficiency of the detector and N is the number of laser pulses summed.

[Palm et al., 2020] calculates the ASR as a function of received surface photons as can be seen in Figure 3.15. As the Figure shows a correlation of 1 between photons received per shot and ASR, it was deemed sufficient to take the number of photons received per cell as a measure of the surface reflectance. As the total distance travelled by an ATLAS laser is variable through a classification cell and the distance travelled is proportional to the amount of laser shots, the number of photons received has been divided by the distance travelled within the cell, to provide the number of photons received per meter. Equation 3.13 provides the calculation for this feature.

$$Photon\ Reflectance = \frac{N_{ph}}{\sqrt{dx^2_{epsg:3035} + dy^2_{epsg:3035}}} \tag{3.13}$$

For a given surface, time, orbit position and laser the returning laser energy is largely effected by clouds and their properties. At the time of writing, this has not been considered in the calculation of this feature. Due to the fact that only classification cells with a sufficient amount of photons are considered, the effect of clouds is considered to be already included in the calculation to a sufficient extent, in other words, classification cells falling under cloudy areas have been removed. Both the absolute and per meter

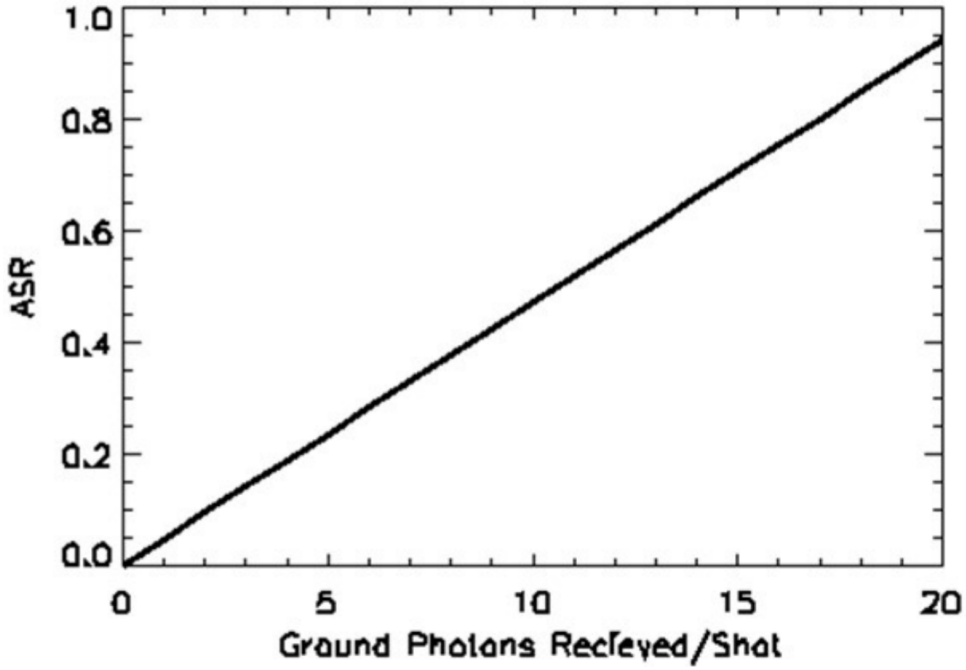


Figure 3.15: The ASR as computed by Equation 3.12 as a function of the number of surface return photons received per laser pulse ($r = 496\ km$, $D_c = 1.1$, $F = 1.0$, $E = 160\ \mu J$, $A_t = 0.43\ m^2$, $S_{ret} = 3.79\ e17$, $N = 400$) [Palm et al., 2020]

values are included as features, called *n_ph*, *n_ph_rel* respectively.

Skewness

The skewness is a measure of the difference in shape or asymmetry of a sample distribution from a bell curve or normal distribution. A positive skew indicates that the tail falls to the left hand of the distribution. The skewness is calculated from the photon heights above the WGS84 ellipsoid of all the photons within the classification cell using Equation 3.14. The x-axis in Figure 3.16 provides the variance in skewness from 1376 samples taken from a laser track across The Netherlands.

The sample skewness is defined as the Fisher-Pearson coefficient of Skewness [Zwillinger and Kokoska, 1999]:

$$Coefficient\ of\ skew = \frac{m_3}{m_2^{\frac{3}{2}}} \tag{3.14}$$

where,

$$m_i = \frac{1}{N_{ph}} \sum_{n=1}^{N_{ph}} (h_n - h_{mean})^i \tag{3.15}$$

Kurtosis

Kurtosis is a measure of the difference of the extent of the tail of a sample distribution with respect to a bell curve or normal distribution. Data sets that have high kurtosis (or heavy tails) tend to have outlier and vice versa for distributions with low kurtosis, where the extreme case would be a uniform distribution. The kurtosis is calculated from the photon heights above the WGS84 ellipsoid of all the photons within the classification cell. The y-axis in Figure 3.16 provides the variance in kurtosis from 1376 samples taken from a laser track across The Netherlands.

The kurtosis of the classification cell is calculated using Equation 3.16 [Zwillinger and Kokoska, 1999]:

$$Coefficient\ of\ Kurtosis = \frac{m_4}{m_2^2} \tag{3.16}$$

where m_i is given by Equation 3.15.

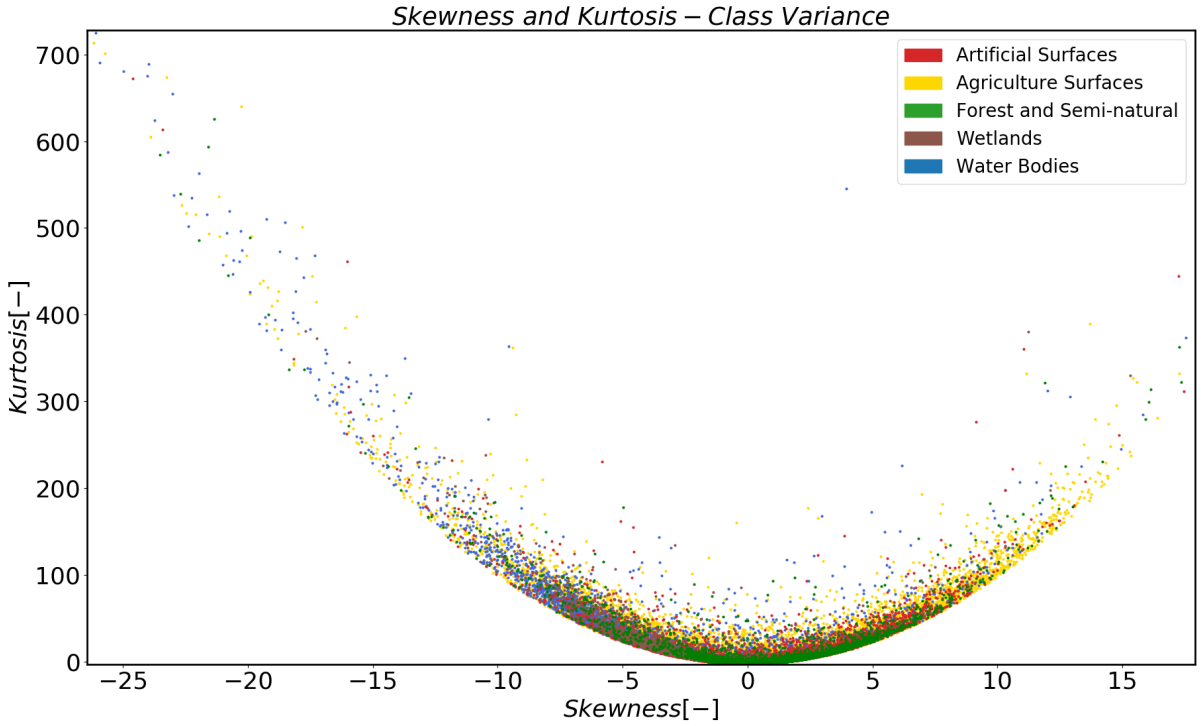


Figure 3.16: Skewness and Kurtosis Variance for all classes. 1376 samples taken from one laser track across The Netherlands

3.5.2 Histogram Features

The next set of features are referred to as "Histogram Features" as they are all derived from histograms created from the heights of the photons above the WGS84 ellipsoid of all the photons in a classification cell. Histograms provide some desirable attributes for this research. Features created from the histograms can be processed much quicker as the data is essentially reduced from the number of photons to the number of histogram bins. In addition, it allows the simple determination of areas of dense photon return, or in other words, the discovery of a scattering surface.

For the analysis presented in this section, a histogram peak is considered analogous to a scattering surface, that is, a defined surface within a classification cell. For example, this could be the roof of a house, the bank of a lake or a dense area of tree canopy. The method of persistent homology was utilised in order to obtain the histogram peaks from each classification cell. Persistent homology is an algebraic method of discerning topographical features of a dataset. In this application, the topographical features of concern are the peaks of the photon height histograms.

The goal with the peak detection algorithm applied is not to just to find the global maximum, as in most cases, this will be the underlying surface. In the case of a lightly forested area, this would negate the less dense canopy and solely detect the underlying surface. Nor is the goal to detect all the local maxima, as many are insignificant compared to other maxima in the neighbourhood. The goal is to identify the relative maxima and quantify their strength with respect to other identified maxima.

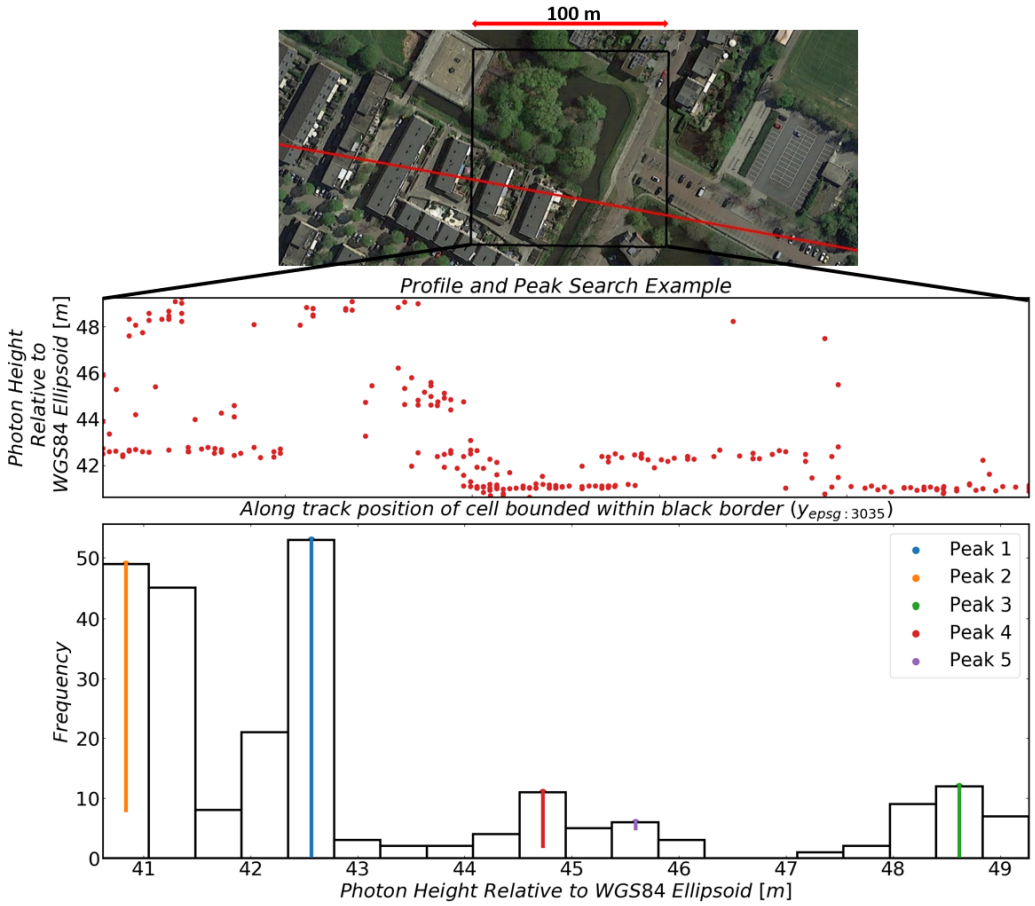


Figure 3.17: Histogram, Profile and peaks identified from peak search algorithm - "Artificial Surfaces"

A mathematical definition of persistent homology is beyond the scope if this report, however Appendix A provides an algorithmic formulation of the problem in order to give some context to the theorem. For the readers comparison, Figures 3.17 to 3.20 present examples of histograms and respective profiles of 4 arbitrary classification cells, from the 'Artificial Surfaces', 'Forest and Semi-natural', 'Wetlands' and 'Water Bodies' classes. A simple analogy to describe the operation of the algorithm is to imagine that the histograms presented in Figures 3.17 and 3.20 are vast mountain ranges lying under the sea in the peak of an inter-glacial period. As the temperature cools the sea level starts to drop, the mountain peaks start to emerge. A peak is created when these local maxima emerge. Whenever a local minimum surfaces, we consider the higher peak to have died, or the lower peak to have merged in to it, which provides a quantity to assess the significance or persistence of the peak. In Figures 3.17 to 3.20, the peaks and respective persistences are depicted by the coloured lines, which are measured from the circle at the top of the coloured line to the bottom of the line. The persistence, which is the photon histogram counts, is considered as a measure of the strength of a scattering surface, as many photons were reflected from this height range.

In Figure 3.17, notice how some of the obvious scattering surfaces are detected by the algorithm. Moving from left to right of the classification cell, the footprint passes over two buildings that seem to be the same height, these are represented by Peak 3. There is the artificial ground surface, represented by Peak 1 and the two bodies of water at the middle and end of the cell, which are lower than the ground surface, represented by Peak 2.

In Figure 3.18, the laser seems to elude the trees until reflecting off a $10m$ tree at the end of the cell, represented by Peak 2. Peak 2, which is much smaller in strength, due to the characteristically low return from vegetation, and appearing at a much higher altitude, will be a strong indicator for a Forest.

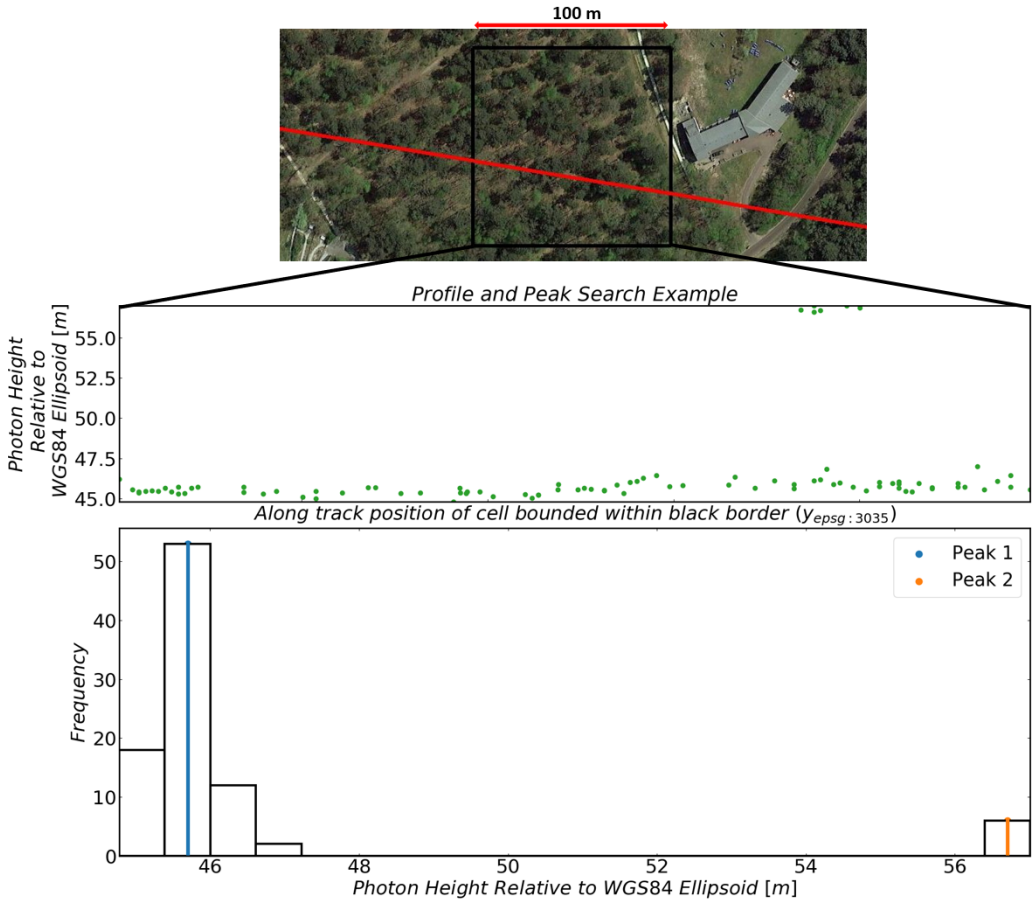


Figure 3.18: Histogram, Profile and peaks identified from peak search algorithm - "Forest"

A Wetlands histogram is provided in Figure 3.19 and shows the characteristically choppy, yet dense profile that is seen from Wetlands. The choppiness comes from the uneven ground associated with vegetation, while the dense photon return comes from the water within the cell. It is interesting to note the skew of this cell in comparison to the Water Bodies histogram, shown in Figure 3.20. The downward facing spikes, typical of Water Bodies, mean that there are tiny peaks below the strongest peak.

It is clear from Figures 3.17 to 3.20 that the strength, altitude and relative position of histogram peaks are class specific. The peaks detected from the histograms of the classification cell can represent valuable information within the cell at a small computational cost. The information provided by the 100's of photons reflected per cell can be reduced to a 20 bin histogram and finally, the location and strength of peaks within it.

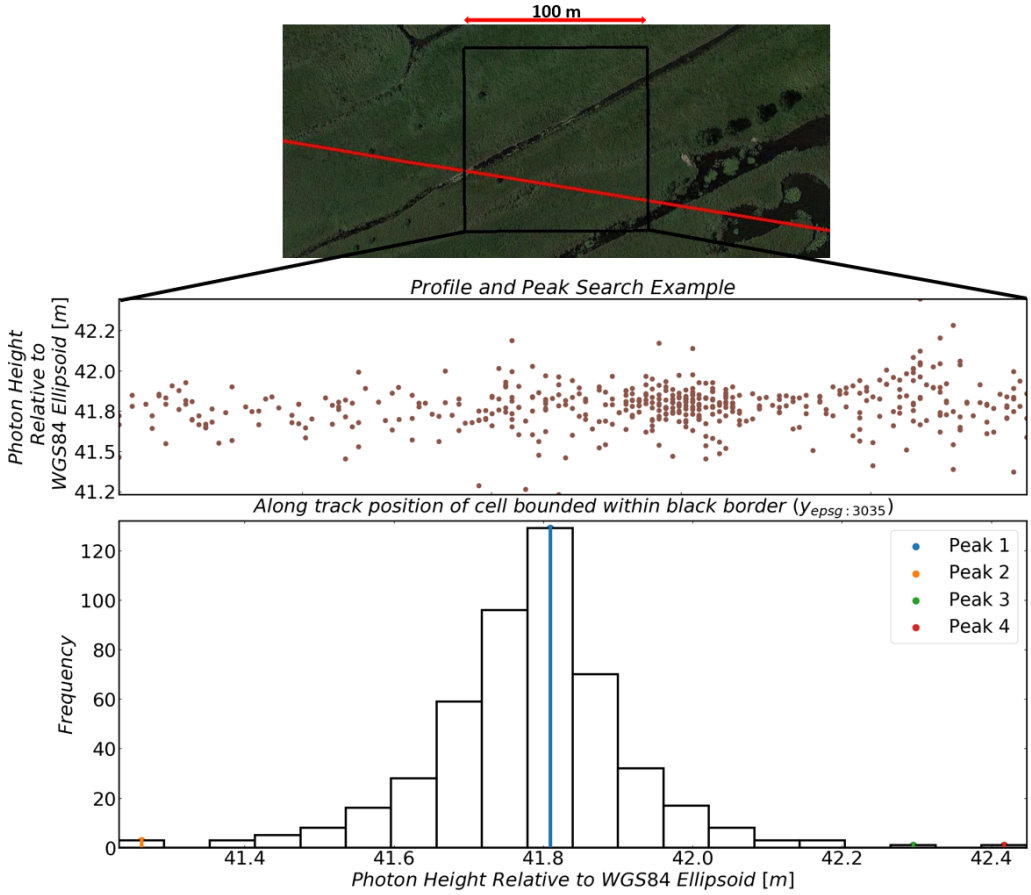


Figure 3.19: Histogram, Profile and peaks identified from peak search algorithm - "Wetlands"

A number of features were calculated from the histogram peaks. For each classification cell, a number of peaks, P_N ($P_N >= 1$) are calculated.

For each classification cell,

$$P = P_1, ..., P_N \tag{3.17}$$

where each P_i has an associated persistence (strength of scattering surface) and location, labelled $P_{i,pers}$ and $P_{i,loc}$ respectively and ordered from the most persistent peak, P_1, to the least persistent P_N. Table 3.1 presents the features that were created from the peaks and their respective persistence's. It would be possible to create many more features from these peaks, given the multiple persistence peaks for a classification cell (dependant on bin size). However, it is intended to keep this list as concise as possible, whilst pertaining the most significant information. The feature values for the histograms in Figures 3.17 to 3.20 are also provided for the readers comparison in Table 3.1. Features are created from the strength

of the strongest, second strongest and weakest persistence's. The heights at which they occur are also features as well as the difference in both height and persistence between them.

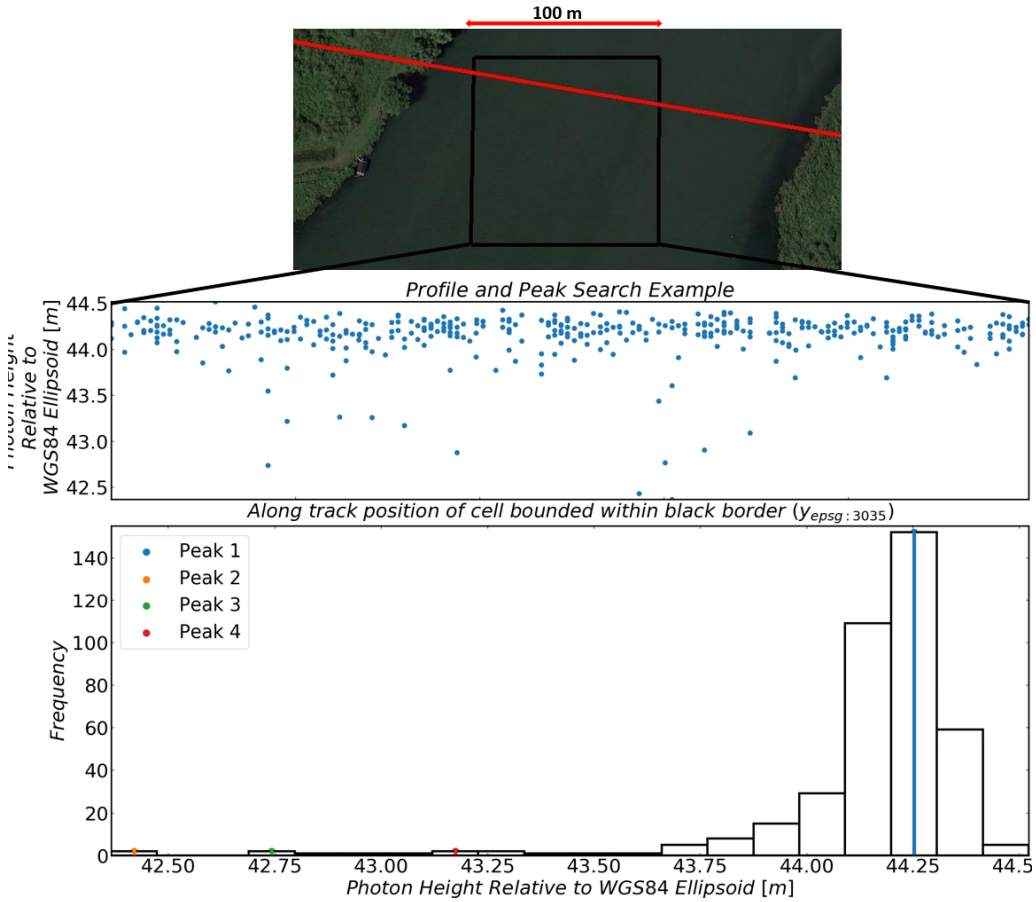


Figure 3.20: Histogram, Profile and peaks identified from peak search algorithm - "Water Bodies"

It is hoped that the features presented in Table 3.1 will assist especially with the discrimination of "Artificial Surfaces", "Agriculture" and "Forest and Semi-natural" as these profiles can be relatively similar, especially in comparison with the "Water Bodies" and "Wetlands" classes, whose profiles are of course alike. In general, it is expected that artificial surfaces will have strong peaks, where photons are reflecting of sharp man-made surfaces such as roofs and roads. In addition, it is likely that these peaks will be further away from each other than in other classes, the transition from a roof to a road for example. In comparison, for the "Forest and Semi-natural class", using a tree as an example, it is expected that the photons will reflect more randomly throughout the canopy of the tree, leading to less defined peaks and less distance between them. In addition to the features defined from the peak detection algorithm, there are some further features calculated from the histogram properties:

Number of Empty Bins
The number of histogram bins with a count of zero. $Empty\ bins = 5$ in Figure 3.17

Maximum consecutive Empty Bins
The maximum number of empty bins that lie next to each other. For example, in Figure 3.17, this would be equal to 4, the consecutive bins appearing from bins 4 and 7.

Feature	**Formula**	**Artificial** *Fig 3.17*	**Forest** *Fig 3.18*	**Wetlands** *Fig 3.19*	**Water Bodies** *Fig 3.20*
N_peaks	$\sum P$	5	2	4	4
Pers_max	$P_{1,pers}$	53	53	129	23
Pers_sec	$P_{2,pers}$	41	6	3	6
Pers_weak	$P_{N_{pers}}$	1	6	1	1
Pers_max_rel	$\frac{P_{1,pers}}{N_{ph}}$	0.218	0.582	0.268	0.137
Pers_sec_rel	$\frac{P_{2,pers}}{N_{ph}}$	0.169	0.066	0.006	0.036
Pers_weak_rel	$\frac{P_{N,pers}}{N_{ph}}$	0.004	0.066	0.002	0.006
dPers_max_sec	$P_{1,pers} - P_{2,pers}$	0.049	0.516	0.261	0.101
dPers_max_weak	$P_{1,pers} - P_{N,pers}$	0.214	0.516	0.266	0.131
Pers_mean	$\frac{\sum_{i=1}^{N_{peak}} P_{i,pers}}{N_{peak}}$	23.2	29.5	33.5	7.75
h_max_peak [m]	$P_{1,loc}$	42.57	45.70	41.81	41.56
h_sec_peak [m]	$P_{2,loc}$	40.84	56.70	41.26	41.21
h_weak_peak [m]	$P_{N,loc}$	45.59	56.70	42.42	40.92
h_max_sec [m]	$P_{1,loc} - P_{2,loc}$	1.72	-11.0	0.55	0.35
h_max_weak [m]	$P_{1,loc} - P_{N,loc}$	-3.02	-11.01	-0.61	0.64
h_peak_mean [m]	$\frac{\sum_{i=1}^{N_{peak}} P_{i,loc}}{N_{peak}}$	44.47	51.20	41.95	41.12

Table 3.1: The 16 features derived from Histogram Peaks and their respective values in the histograms presented in Figures 3.17 to 3.20

3.5.3 Eigenvalue-based Features

ATLAS data differs from traditional laser scanning data by only providing dense photon returns in two dimensions, that is, in the along track and height directions. Technically, the data does have 3 dimensional variation, however the variance in the across track direction is an order of magnitude smaller than the along track and height directions. It is stated in [NSIDC, 2019] that the footprint of ATLAS is $14m$, however what this means is a little unclear. It is expected that this means the photons can be returned from within a $14m$ corridor parallel to the along track direction of the laser. Figures 3.21 and 3.22 provide 3D views of real segments of the laser tracks above The Netherlands.

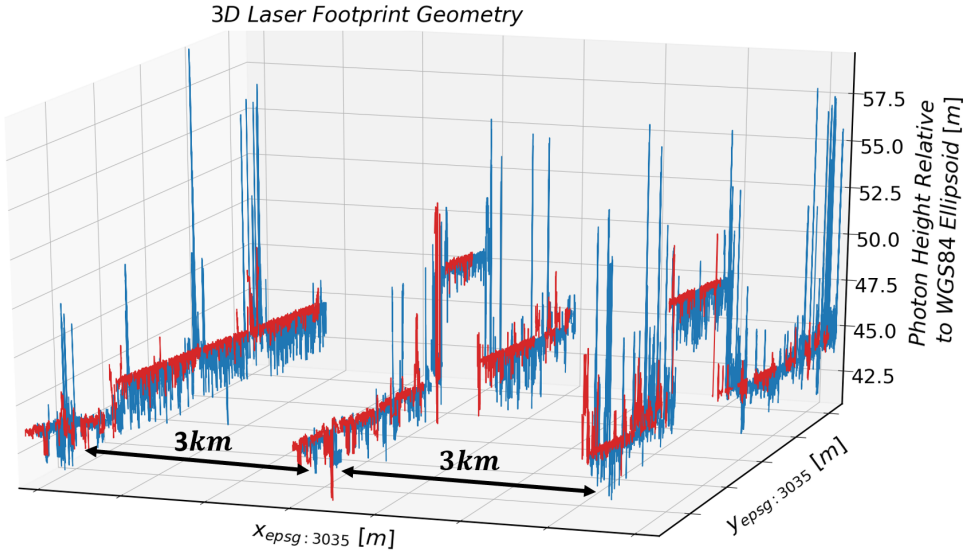


Figure 3.21: Side view of a segment of laser tracks in The Netherlands. x-axis = $7.78km$, y-axis = $4.91km$ Distance between pairs of lasers = $2.5km$

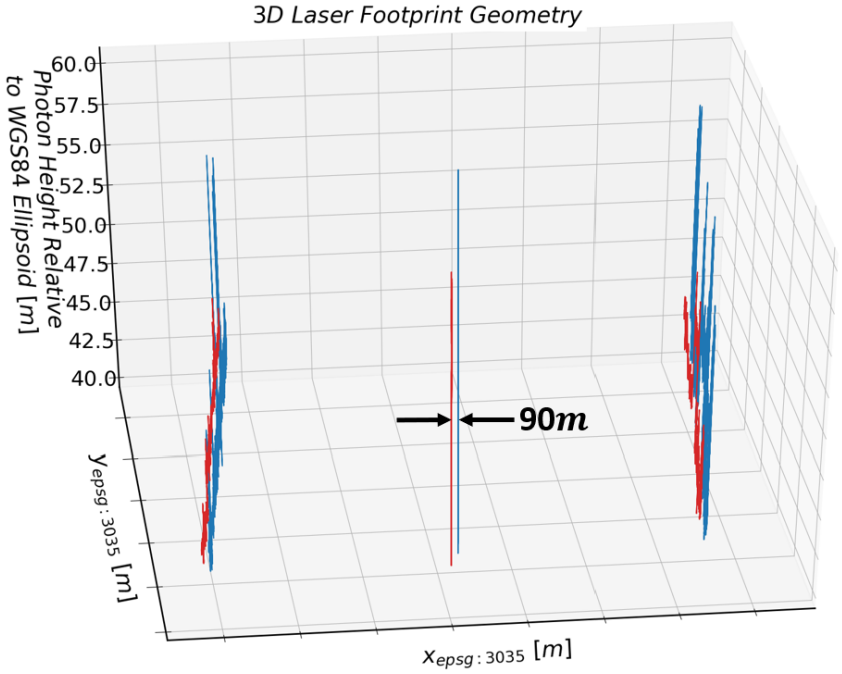


Figure 3.22: Front View of a segment of laser tracks in The Netherlands. x-axis = $7.78km$, y-axis = $4.91km$, Distance between lasers = $90m$

Figures 3.21 and 3.22 provide an approximately $5km$ segment of the laser tracks for all 6 beams. The general geometry can be seen, that is, the $3km$ between the pairs of lasers and $90m$ between the lasers themselves. It is clear that over this $5km$, the lasers don't diverge much from their path, or at lease it cannot be made out from Figure 3.22. The across track variation must be much smaller than $90m$. In order to gauge the across track variation of a single laser footprint print, a linear line of best fit has been plotted of all the photons received from an arbitrary laser track across The Netherlands, this is depicted by the blue plot, with the photons represented by the scattered blue dots. The average deviation from this line on either side is represented by the red and purple plots and the extreme deviation by the orange and green plots. The average deviation from this line is $\approx 2.8m$, with the maximum at $\approx 11.7m$. In other words, it seems that there is a corridor of up to $23.4m$ within which a photon could fall. With most of the information contained within a $5.4m$ corridor. This is larger than the $14m$ previously stated, however it is probably due to the fact that photons arriving from another source are reflected into the instrument. It is clear that the information stored in this dimension is less significant than the other dimensions, however, it has been decided to calculate some basic 3D features for the classification testing. This functionality is implemented as it is desired that the method provided in this report can be applied to classification cells at a higher resolution. If one were to classify at a resolution of $5m$ classification cells or less, it would not make sense to ignore this dimension, given that it's variation can be up to $23m$. In addition, the features chosen have respected the aim to stay computationally fast.

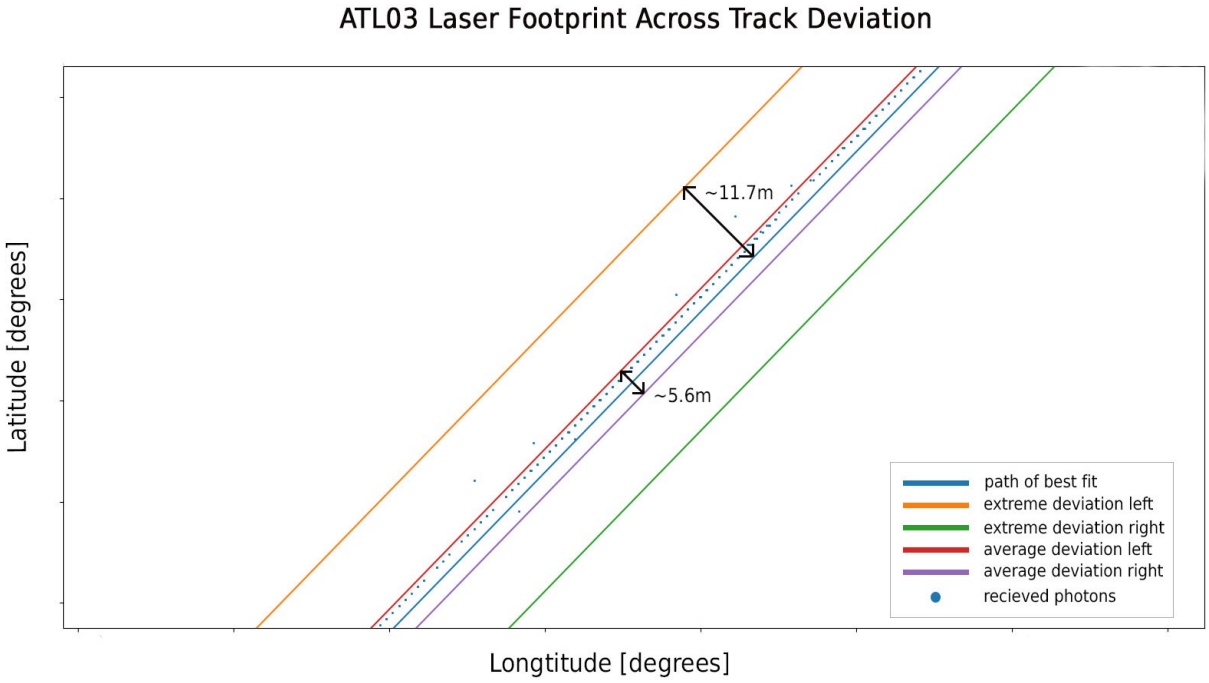


Figure 3.23: Cross-Track Visualisation

The 11 3D features presented here are taken from [Weinmann et al., 2015] and are all geometric features. For each classification cell, let $X_i = (x_i, y_i, z_i)^T$ and $\bar{X} = \frac{1}{n}\sum_{i=1,n} X_i$, the centroid of the n photons. Given $M = (X_1 - \bar{X}, ..., X_n - \bar{X})^T$, the 3D structure tensor is defined by the covariance matrix $C = \frac{1}{n}M^T M$. An eigenvalue decomposition of the covariance matrix provides the 3 principle directions (eigenvectors) of the photons within a classification cell and their magnitudes (eigenvalues). Figure 3.24 provides the principle directions for 100 random classification cells of each class. The principle directions can be considered a decomposition of the directional variance of a classification cell. There is a clear difference in variance between the 3 principle directions. In terms of classification, this indicates the discriminative potential of a feature created from the eigenvalues of a certain principle direction. Principle direction 1 seems to have the least variance as well as no clear distinction between classes. Principle direction 2 has the largest variance and seems to have more of a distinction between classes, although a pattern is not clear. Saying that, there seems to be the most variation in classes that contain

the most varied surface types, such as Forest, whose vectors are pointing in many directions, as oppose to Water Bodies, whose vectors seem to be pointing in two distinct directions, namely, along the positive and negative x-axis, albeit with some anomalies. Principle direction 3, whose values are mostly positive as it is associated with the photon heights (whose values are all positive), discriminates the most between classes. Water Bodies, Forest and Wetlands can be seen grouped together.

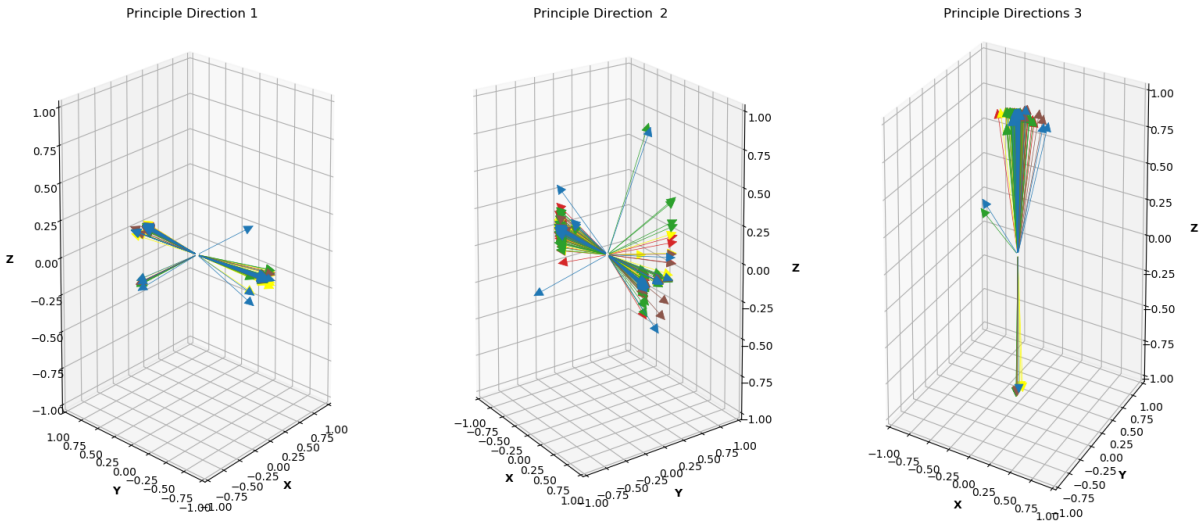


Figure 3.24: Principle Directions for 500 classification cells chosen at random (100 for each class)
Key: **Artificial Surfaces**, **Agriculture**, **Forest**, **Wetlands**, **Water Bodies**

Using the eigenvalues derived from the structure tensor a collection of "Eigenvalue-derived" features were created. Each classification cell is considered as its own cloud, with the photons within it being used as the points. The following features are calculated from the normalised eigenvalues, e_i, where $i \, \epsilon \, \{1, 2, 3\}$, of the photons heights extracted from a classification cell.

$$Linearity = L_\lambda = \frac{e_1 - e_2}{e_1} \tag{3.18}$$

$$Planarity = P_\lambda = \frac{e_2 - e_3}{e_1} \tag{3.19}$$

$$Scattering = S_\lambda = \frac{e_3}{e_1} \tag{3.20}$$

$$Omnivariance = O_\lambda = \sqrt[3]{e_1 e_2 e_3} \tag{3.21}$$

$$Anisotropy = A_\lambda = \frac{e_1 - e_3}{e_1} \tag{3.22}$$

$$Eigentropy = E_\lambda = -\sum_{i=1}^{3} e_i ln(e_i) \tag{3.23}$$

$$Change\ of\ Curvature = C_\lambda = \frac{e_3}{e_1 + e_2 + e_3} \tag{3.24}$$

$$Verticality = 1 - n_z \tag{3.25}$$

where n_z is the vertical component of the normal vector $\mathbf{n} \, \epsilon \, R^3$

$$sum\ EVs_{3D} = (e_1 + e_2 + e_3)_{3D} \tag{3.26}$$

[Weinmann et al., 2015] discusses how the methods applied here are intended to be de-coupled from the size of the neighbourhood used. Still, this method is generally applied to denser photon clouds over much smaller areas. For example, this method is widely used in the object segmentation of LIDAR data. While it has been shown that the variation in the z-direction is much larger than the other two dimensions, it is still thought that there could be valuable information stored within them and it has therefore been decided to use the eigenvalue-based features in the analysis.

3.6 Classification and Test Set-up

In Section 2.3, a relatively high level introduction of the random forest technique was provided, along with a discussion of the most important hyper-parameters/attributes. This Section will provide a motivation for the attributes chosen for the land type classification presented in this report. Sections 3.6.1 and 3.6.2 discuss the hyper parameter and sample distribution selection and Section 3.6.3 discusses the One vs All classification technique used to obtain the most significant features for the classification. Section 3.6.4 provides the final test-up used in the classification. Table 3.4 provides an overview of the features used in the classification and Table 3.3 provides the chosen parameters used as inputs in the classification model.

3.6.1 Hyper-Parameter Selection

The following section will present some tests that were performed in order to motivate the decisions that were made with regards to the hyper-parameters and input sample distribution. Of course, an optimal solution is desired, however, the goal of the research is not to optimise the classification accuracy and therefore significant effort was not spent in perfecting these parameters. An introduction to the random forest classification model and associated parameters can be found in Section 2.3. The tests presented in this Section were performed on the training data, which consisted of 95,562 classification cells and for each test the average accuracy is taken from a 5-fold cross validation.

For the amount of trees used in the forest, in contrast to [Criminisi et al., 2011], where the monotonic increase in test accuracy with an increase in forest size was discussed, Figure 3.25 shows that the validation accuracy drops off after 500 trees. It may be that the test accuracy (performance on unseen data) does indeed monotonically increase with a larger forest. It is decided to use a forest with 500 trees, which provided the maximum validation accuracy with an acceptable computation time.

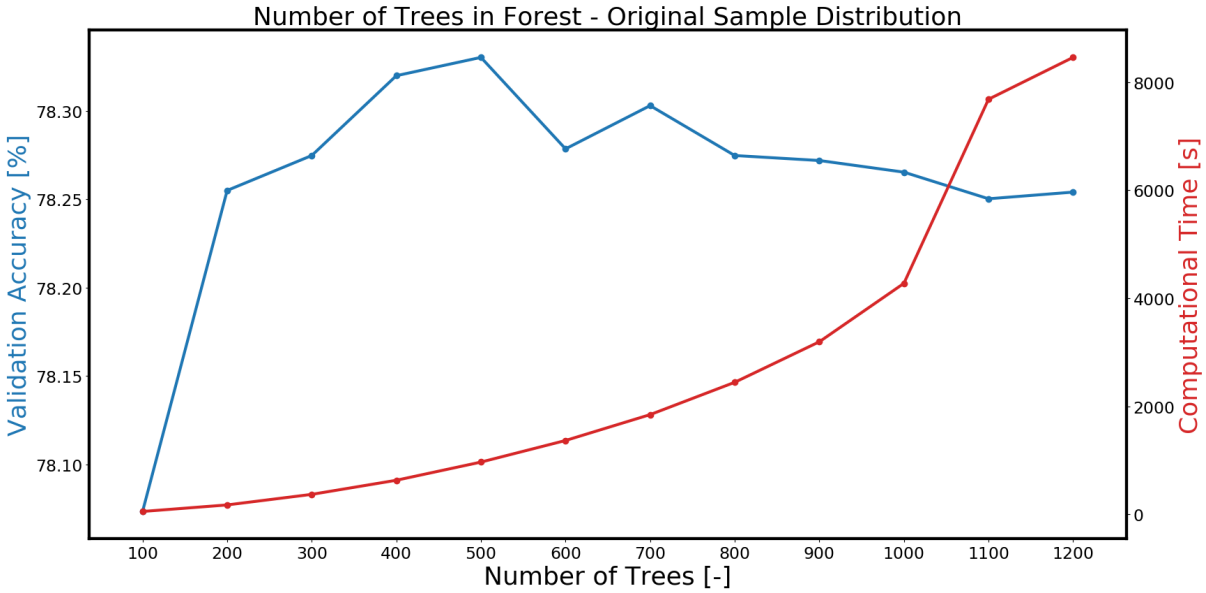


Figure 3.25: Validation accuracy for an increasing Forest Size.
Test Parameters: Tree Depth = 20, Max Features = $\sqrt{N}_{features}$, $N_{samples}$ = 95,562 with 5-fold cross validation

Figures 3.26 and 3.27 provide the validation accuracy for an increasing tree depth for two different sample distributions, that is, the original sample distribution and a randomly undersampled subset of the original distribution. The effect of a sample distribution is discussed in more detail in Section 3.6.2 and the Figures are provided here as a means of comparing the hyper parameters. In both Figures, it can be seen that the validation accuracy jumps up sharply after a depth of 20 is reached and then either stays

approximately the same or changes by a small amount ($\leq \frac{2}{10}\%$) as the tree depth increases, before converging. As discussed in Section 2.3.2, a higher tree depth can lead to overfitting. It is therefore decided to use 30 as the maximum tree depth, as this provides close to the highest accuracy in most of the cases presented in Figures 3.26 and 3.27. It is hoped that this tree depth, will allow for good performance while allowing the model to keep a sufficient level of generalisation, that is, effectively classify unseen data.

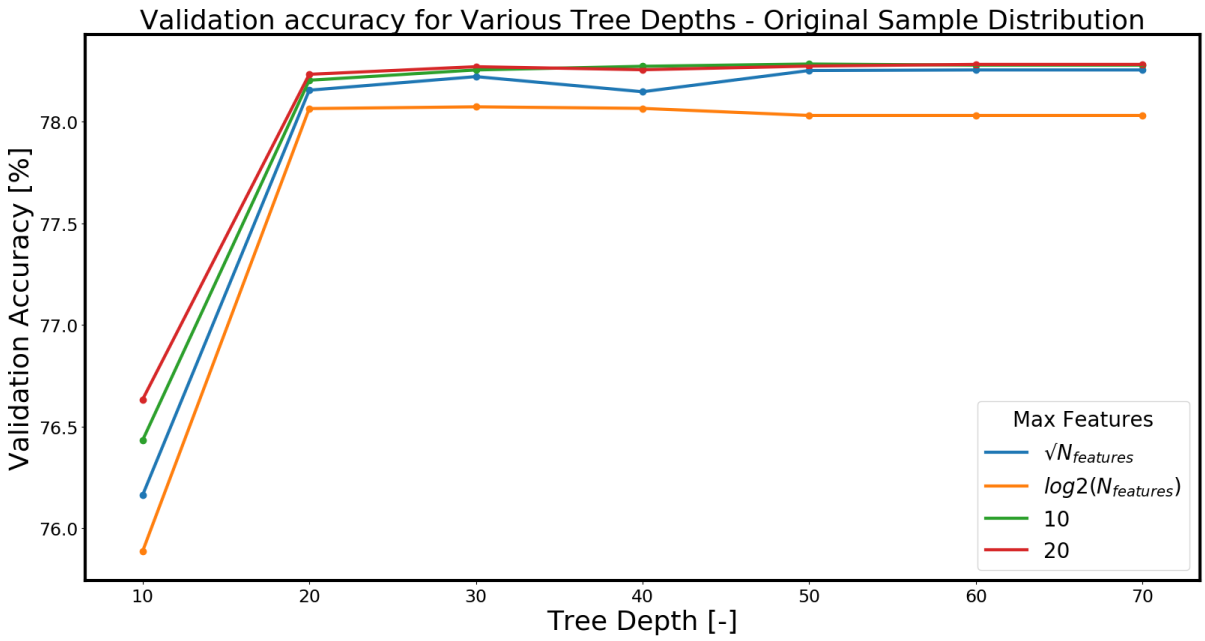


Figure 3.26: Validation Accuracy for an increasing tree depth and Max Features - Original Sample Distribution
Test Parameters: Number of Trees = 100, $N_{samples}$ = 95,562 with 5-fold cross validation

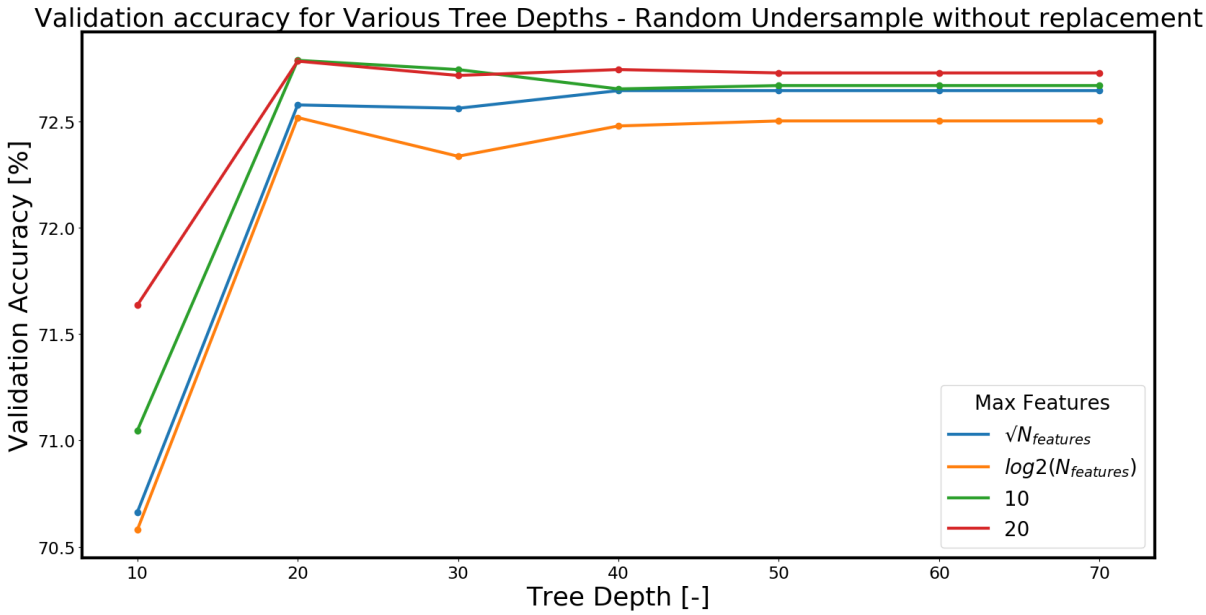


Figure 3.27: Validation Accuracy for an increasing tree depth and Max Features - Randomly Undersampled Distribution
Test Parameters: Number of Trees = 100, $N_{samples}$ = 22805 with 5-fold cross validation

The different coloured plots in Figures 3.26 and 3.27 also indicate the "Max Features" attribute. This is discussed in Section 2.3.2 and represents the amount of features that are considered at each test function/split. The lower this number, the less correlated the trees are with one another. Similarly to the tree depth, there is a only a fraction of a percent difference in validation accuracies between the options tested. It was decided to use $log2(N_{features})$ as it achieved close to the highest accuracy, while being the lowest value of max features, which maximises the amount of randomness injected into the forest.

3.6.2 Sample Distribution

Figure 3.28 shows the sample distribution over the 5 land types considered in this report. Each sample represents a 100x100m classification cell. The dataset is quite heavily imbalanced, with Agriculture having significantly more samples than the other classes and Forest and Wetlands significantly less.

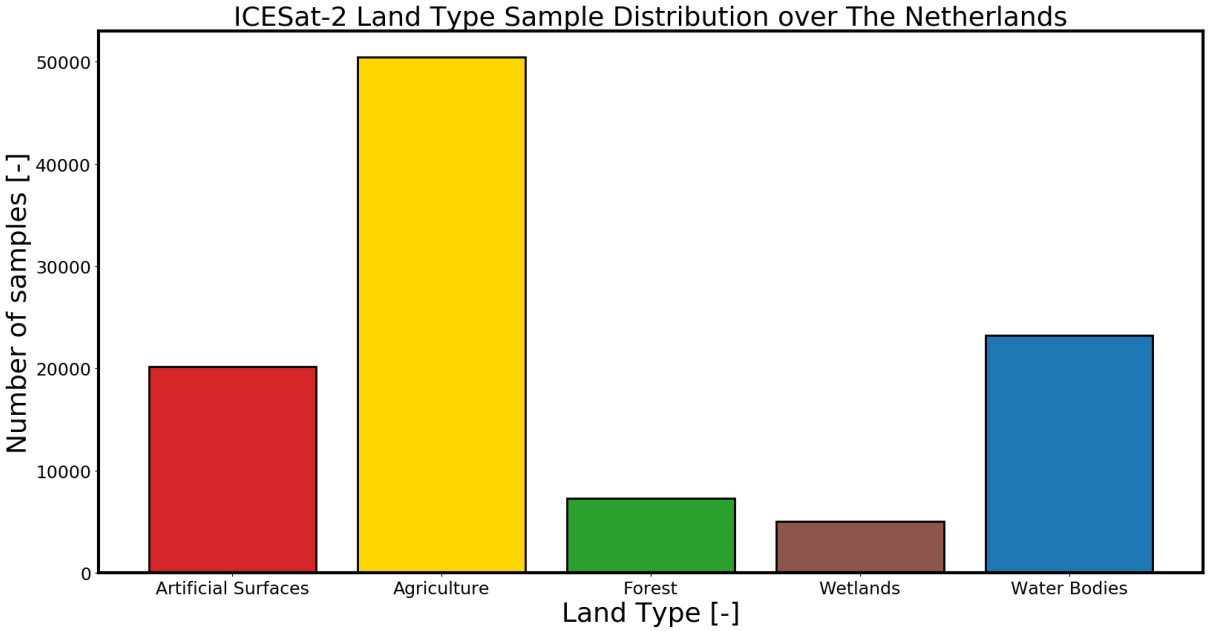


Figure 3.28: Original Sample Distribution of ICESat-2 - Cycle 5 data. Acquired between September 26^{th} - December 26^{th} 2019, $N_{samples} = 106181$

It is discussed in [Criminisi et al., 2011] that an unbalanced class distribution can be detrimental to forest performance. To this end, a variety of different input sample distributions were investigated. Table 3.2 provides the accuracy results per class for the different distributions tested. The simplest way to overcome a class imbalance is to re-sample the data set to create a uniform or close to uniform (stratified) dataset. Artificially making the class balance equal by down-sampling the majority class or over-sampling the minority class has been shown to be more effective than using a largely imbalanced dataset [Drummond et al., 2003]. In Table 3.2, "Randomly Undersampling - min samples" refers to randomly selecting samples from all the classes equal to the amount of samples of the lowest class, in this case, an equal amount of samples as the Wetlands class. The replacement refers to whether or not a chosen sample was returned to the pool before selecting the next sample. "Random Oversampling - Max samples" is the opposite, where samples are randomly selected from the pool until there is an equal amount as the majority class. Of course, replacement is necessary here. An additional technique for oversampling is the Synthetic Minority Over-sampling Technique (SMOTE), which was created in order to combat the common problem that a certain class is largely underrepresented in a dataset. The technique is executed by taking each majority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbours [Chawla et al., 2002]. The final 3 sampling distributions shown in Table 3.2 were a combination of undersampling and SMOTE oversampling in order to create an equal class balance.

Sampling Technique	**Overall Accuracy [%]**	Artificial [%]	Agriculture [%]	Forest [%]	Wetlands [%]	Water Bodies [%]
Original Distribution	76	42	90	28	60	79
Random Undersampling (min samples, with replacement)	75	72	62	63	90	85
Random Undersampling (min samples, without replacement)	67	61	57	61	80	78
Random Oversampling (Max Samples)	97	99	85	100	100	99
SMOTE Oversampling (Max Samples)	80	76	76	82	86	82
Combined SMOTE/ Under sampling (Equal Classes/ 30000 per class)	74	64	65	75	84	81
Combined SMOTE/ Under sampling (Equal Classes/ 50000 per class)	76	70	71	78	84	80
Combined SMOTE/ Under sampling (Equal Classes/ 70000 per class)	79	64	65	75	84	81

Table 3.2: Sample Distribution Testing - Per class validation accuracy

The first important conclusion drawn from Table 3.2 is the high sensitivity of the model to the sample distribution. The validation accuracy, however, does in general remain at an acceptable level above 70%. The overall accuracy of the "Original Distribution", presented in Figure 3.28, is misleading at 76% as there was a large imbalance in the per class performance. Unsurprisingly, Agriculture, the heavily majority sampled class achieved a superior validation accuracy (90%) in comparison to the lesser sampled Forest class (28%) for this distribution. "Random Undersampling - with replacement" was able to achieve almost the same performance (75%) but with a more even performance across the classes. However, the problem with using replacement is that it is highly likely that there are samples in the test data that have been seen in the training data. This is undesirable as it reduces your models ability to successfully classify unseen data and provides a misleading performance indication. In addition, in the act of resampling, valuable information is already being lost in the omission of data samples and replacement further contributes to this. Indeed, "Random Undersampling - without replacement" has an inferior validation accuracy at 67%, due to the fact that is is only being tested on data it hasn't seen before. "Random Oversampling - Max samples" achieved a near perfect classification, but of course, has likely seen almost all of the test samples within the training phase, where the Wetlands samples would all have been resampled more than 10 times. SMOTE would be the better option if one were to oversample as it creates unique datapoints whilst achieving a high performance. The final three distributions are a combination of undersampling and SMOTE, producing an increasing performance for an increasing amount of samples per class.

The main conclusion that was taken from the testing of different sample distributions was the fact that the model is significantly sensitive to the sample distribution. In fact, determining the optimal distribution for this application could be a study of its own and dealing with imbalanced datasets is indeed hot topic in machine learning. As a result of the discovered sensitivity to sample distribution, it was decided to not go further than what was presented in Table 3.2. In lieu of investigating the optimal sample distribution strategy, it has been attempted to reduce the effect of the sample distribution choice on the final result. It was therefore decided to move forward with "Random Undersample - without replacement", where the number of samples per class equalled that of the minority class. It may seem unintuitive as this was the distribution that achieved the lowest classification accuracy, however, considering the aims of the research, to determine whether the raw data and its extracted features offer classification potential, it seems the sensible option. By creating an equal set of unique samples for each class, the overall relative performance of each class can best be evaluated, as well as the features that are most important at classifying them. The introduction of synthetic examples and/or repeat samples may skew the result and reduce the ability to observe which features best contribute to the classification of each class. The downside of this strategy is that much of the training data is essentially thrown away.

3.6.3 Feature Importance and the One vs All Classification Strategy

In order to determine which features are most important for the classification, the gini-importance measure was used for each split function in each tree of the forest. As introduced in Section 2.3.1, the gini-impurity (Equation 2.3) measures the ability of a feature (or node test function) to discriminate between classes, that is, how effectively can the feature split up the samples. The feature importance of a decision tree is calculated by summing up the gini-importance for the feature used at each test function and the feature importance over the whole forest is the mean of the importance calculated for each feature in every tree. Once the classification has been performed, a relative performance of each feature in the classification can be obtained.

As a means to address the research question "Which features unique to ATL03 photon data are best at distinguishing land types from one another?", the One vs All classification technique is employed. This turns the multi-class problem that was used for the overall classification into a binary problem [Huang and Fisher, 2014], where the class in question is classified against the other four classes. In this way, the class under scrutiny is better isolated and the features that discriminate it from the rest can be better observed. For the feature importance testing with use of the One vs All classification strategy, a separate model is trained using the same train and test data as with the general classification model and all the same hyper-parameters are used as well as a "Randomly undersampled - without replacement" sample distribution. The only difference is that there are only two outcomes that the forest can predict, namely, whether it is the class under scrutiny or not, the not meaning it belongs to any of the other four classes.

3.6.4 Test Set-up

This section presents the final test set-up for the classification model. The data used for the classification was taken from ICESat-2's 5th repeat orbit, Cycle 5 and is acquired between September 26^{th} 2019 and December 26^{th} 2019. It is expected that the effect of seasons may play a role on the quality of the classification or at least have an effect on the characteristic profiles. For example, it is expected that a forest profile will look different depending on whether or not the forest has leaves on its trees. Due to this, it was decided to perform the test on data taken from the same cycle/season, such that the effect of a changing climate on the results is minimised. Indeed, this would be an interesting area to investigate.

Data that was too cloudy or reflective was omitted, that is, classification cells that didn't return the minimum number of photons deemed acceptable for classification. This left $106,181$ classification cells, which were split up into 90% training data and 10% testing data. The testing data was taken out at the start of the project and was not used for any of the model/feature development. As the test data is taken out before any re-sampling is performed, it adheres to the true sample distribution, that is, approximately the same balance as that presented in Figure 3.28. 10,619 samples were used for testing.

An overview of the 51 features used in the classification, which were introduced in Section 3.5, are provided in Table 3.4 and an overview of the hyper-parameters used in the classification algorithm are provided in Table 3.3. These tables hold for both the general classification and the One vs All classification used to obtain the feature importance. The "scikit-learn" [Pedregosa et al., 2011] toolbox was utilised in the creation of the classification algorithm, where the "RandomForestClassifier" [scikit-learn, 2020] provided most of the Random Forest algorithm functionality.

Attribute	**Value**
Number of Trees	400
Maximum Tree Depth	30
Maximum Features	$log2(N_{features})$
Sampling Strategy	Randomly Undersampled without replacement
Bagging/ Bootstrapping	Yes
Split Criteria	Maximise Gini-Impurity
Number of Training/Test Samples	95562 (90%) / 10619 (10%)

Table 3.3: Random Forest Classification Attributes

Feature	Unit	Description
h_mean	[m]	Mean height of photons
N_ph	[-]	Number of photons received
h_range	[m]	Height range of photons
h_std	[m]	Standard Deviation of photon heights
h_skew	[-]	Skew of photon heights
h_kurt	[-]	Kurtosis of photon heights
N_ph_rel	$[1/m]$	Number of photons recieved per meter
h_var	[-]	Variance of photon heights
h_normal	[m]	Height difference between mean & minimum height
h_min	[m]	Minimum photon height
h_max	[m]	Maximum Photon height
HRR	[-]	Ratio of photons above and below half of photon range
fifth_perc	[m]	5th percentile of the photon heights
ten_perc	[m]	10th percentile of the photon heights
quart_perc	[m]	25th percentile of the photon heights
half_perc	[m]	50th percentile of the photon heights
three_quart_perc	[m]	75th percentile of the photon heights
ninety_perc	[m]	90th percentile of the photon heights
ninety_fifth_perc	[m]	95th percentile of the photon heights
perc_95_5	[m]	Height difference between the 95th & 5th percentiles
perc_75_25	[m]	Height difference between the 75th & 25th percentiles
perc_50_5	[m]	Height difference between the 50th & 5th percentiles
Linearity	[-]	Linearity of photon heights
Planarity	[-]	Planarity of photon heights
Scattering	[-]	Scattering of photon heights
Omnivariance	[-]	Omnivariance of photon heights
Anisotropy	[-]	Anisotropy of photon heights
Eigentropy	[-]	Eigentropy of photon heights
Change_of_curvature	[-]	Change of curvature of photon heights
Verticality	[-]	Verticality of photon heights
sum_EVs_2D_xz	[-]	The sum of the 2D eigenvalues in the x & z plane
sum_EVs_3D	[-]	The sum of the 3D eigenvalues
EV_ratio_2D_xz	[-]	The ratio of the eigenvalues in the x & z planes
N_peaks	[-]	The number of histogram peaks detected
Pers_max	[-]	The maximum persistence of the histogram peaks
Pers_max_rel	[-]	Maximum persistence divided by number of photons
Pers_sec	[-]	Persistence of the second most persistent histogram peak
Pers_sec_rel	[-]	Second most persistent peak divided by number of photons
Pers_weak	[-]	Persistence of the least persistent histogram peak
Pers_weak_rel	[-]	Least persistent peak divided by number of photons
dPers_max_weak	[-]	Persistence difference of most & least persistent peaks
dPers_max_sec	[-]	Persistence difference between maximum & second persistent peaks
Pers_mean	[-]	Average persistence of the histogram peaks
h_max_peak	[m]	Height of the most persistence peak
h_weak_peak	[m]	Height of the least persistent peak
h_sec_peak	[m]	Height of the second most persistent peak
h_peak_mean	[m]	Average height of the histogram peaks
h_max_weak	[m]	Height difference between most & least persistent peaks
h_max_sec	[m]	Height difference between maximum & second persistent peaks
bins_empty	[-]	Number of empty histogram bins
bins_empty_consec	[-]	Maximum number of consecutive histogram bins

Table 3.4: List of 51 Features used in the classification grouped by feature type: *22 Altitude Derived Features* - Section 3.5.1 ,*18 Histogram-based Features* - Section 3.5.2, *11 Eigenvalue-based Features* - Section 3.5.3

4. Classification Results & Interpretation

This Section provides the classification results for the experimental set-up outlined in Section 3. The classification strategy and test set-up can be found Section 3.6, while a discussion of the features created for the classification can be found in Section 3.5.

The results are separated into 3 sections, which all address a different aspect of the classification. Section 4.1 provides the overall results of the classification in The Netherlands. Section 4.2 provides the results of the One vs All classification strategy which aimed to shed light on which features where most important at distinguishing one class from another. Finally, Section 4.3 provides a misclassification analysis performed on a subsample of the misclassified classification cells. The aim here was to categorise and quantify error types, that is, reasons for predictive confusion. This Section also addresses the suitability of the CORINE data set for the classification of ATLAS data.

4.1 Overall Accuracy

The overall accuracy of the model considers the predictions of the test data fed into the classification model. When the predicted class output from the algorithm was the same as the CORINE class, the results was considered positive. As described in Section 3.6, there were 10,619 test samples. The sample distribution of the test data was unchanged, as oppose to the training data, which was re-sampled such that each class had the same amount of samples. A motivation for the selection of sample distribution can be found in Section 3.6.2. In order to reduce bias toward a particular weather condition or data acquisition time, the test samples were selected randomly from all applicable samples across the three month test period. This means that the test samples have been distributed evenly across The Netherlands according to the sample class density of the acquisition data.

The classification result can be seen in Figure 4.1. Each coloured pixel represents the predicted class of the respective $100m$ classification cell. The $100m$ classification pixels have been enlarged in this Figure for easier observation. The along track gaps in the tracks indicate the minimum number of photons were not returned for a particular classification cell, due to weather/daylight conditions or poor reflectivity. For all of the tracks, the path of the track can be identified and of course the areas in between cannot be classified, as no data was acquired there. It is interesting to note the general trends across the country. As shown in the original data distribution and test sample distribution, agriculture largely outweighs all other classes. The Randstad can be observed in the South to Mid-West of the country by the relatively dense pockets of Artificial Surfaces, between Breda and Amsterdam. The artificial lakes IJsselmeer and Markemeer can be seen in the middle North of the country by the dense area of Water Bodies and Wetlands above Lelystad. Forests and Semi-natural areas can be seen dotted throughout the country, with the most dense area lying through the "Hoge Veluwe" national park between Almere and Apeldoorn as well as at the Belgian border.

Figure 4.2 and 4.3 provide further examples of the classification result. The predicted test classification cells have been superimposed over Landsat/Copcernicus observation satellite images provided by Google Earth. As can be seen in Figure 4.1, the density of test samples varies across the country and the examples provided in Figures 4.2 and 4.3 provide areas with a relatively high density of test samples. Figure 4.2 shows the return over the port of Antwerp at the Dutch-Belgian border. In the left of the image, a mixture of Artificial surfaces and Water Bodies can be seen as the tracks pass over the winding port area. Leading up to this from the right, the transition from countryside can be seen. Figure 4.3 provides the classification above Texel, in in the North-West of The Netherlands. This provides a good example of the distinction between Wetlands and Water Bodies, where the algorithm seems to have successfully detected the inter-tidal flats that constituent the Wetlands class, indicated by the brown classification cells. To the left of the image, the sandy island tip of Vlieland has been successfully classified as Forest and Semi-natural areas, a positive result considering the limited training samples for beaches in comparison to the other constituents of the Forest and Semi-natural areas class.

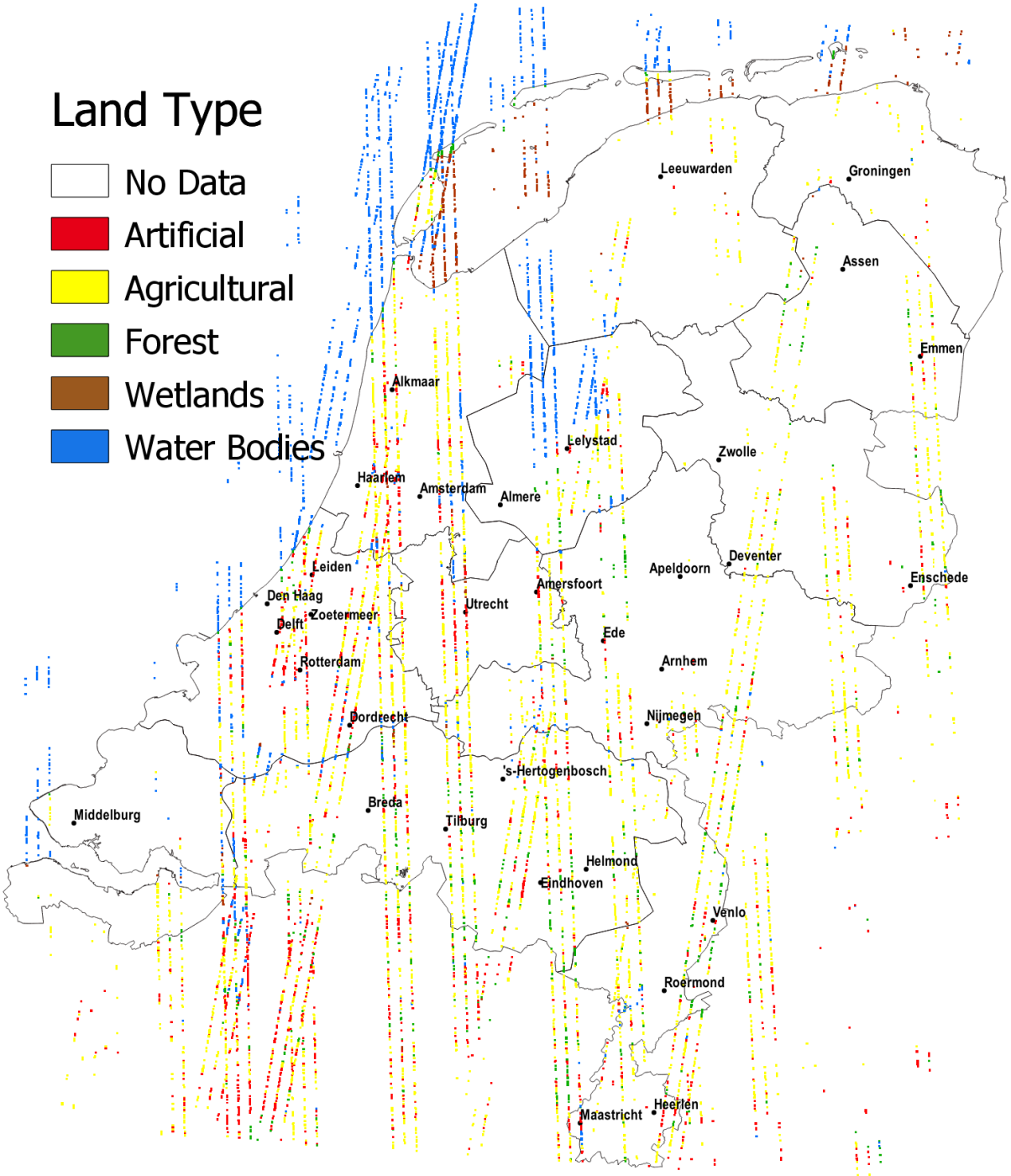


Figure 4.1: Classification Result - The Netherlands. Each pixel represents the predicted class for a $100m$ classification cell. Pixels have been enlarged by an order of 3 for easier observation.

Figure 4.2: Classification Result - Texel and Den Helder
Key: **Artificial Surfaces**, **Agriculture**, **Forest**, **Wetlands**, **Water Bodies**

Figure 4.3: Classification Result - The Port of Antwerp
Key: **Artificial Surfaces**, **Agriculture**, **Forest**, **Wetlands**, **Water Bodies**

The overall classification accuracy of the 10,619 test samples was 71.2%. The detailed classification results are presented in the confusion matrix in Figure 4.4. The confusion matrix provides per class information on the true and predicted classes. For example, with the origin in the upper left, cell (1,2) indicates that 276 samples classified as Artificial Surfaces by the validation data, were deemed to be Agricultural areas by the classification algorithm. The diagonal of the confusion matrix signifies the successful results. The best performing classes are Wetlands and Water Bodies with classification accuracies of 81.8% and 84.2% respectively. This was to be expected due to the low variability of surface types within the class. For example, referring to the class surface breakdown in Figure 3.5, Water Bodies consists of only Inland Waters and Marine Waters. The only significant feature separating these two classes would perhaps be changing surface height of ocean water due to the presence of waves. Water profiles have distinct characteristics, such as the return of photons from below the water surface (see Figure 3.3d) as well as characteristically low variability with respect to the Artificial Surfaces, Agricultural Areas and Forest and Semi-natural areas classes. For instance, the surface profiles of water classes would generally be flat, with a distinctly low average height and range in comparison to other classes. Naturally, a large proportion of the confusions for the water classes are between each other, 5.3% and 8.7% for Wetlands and Water Bodies respectively.

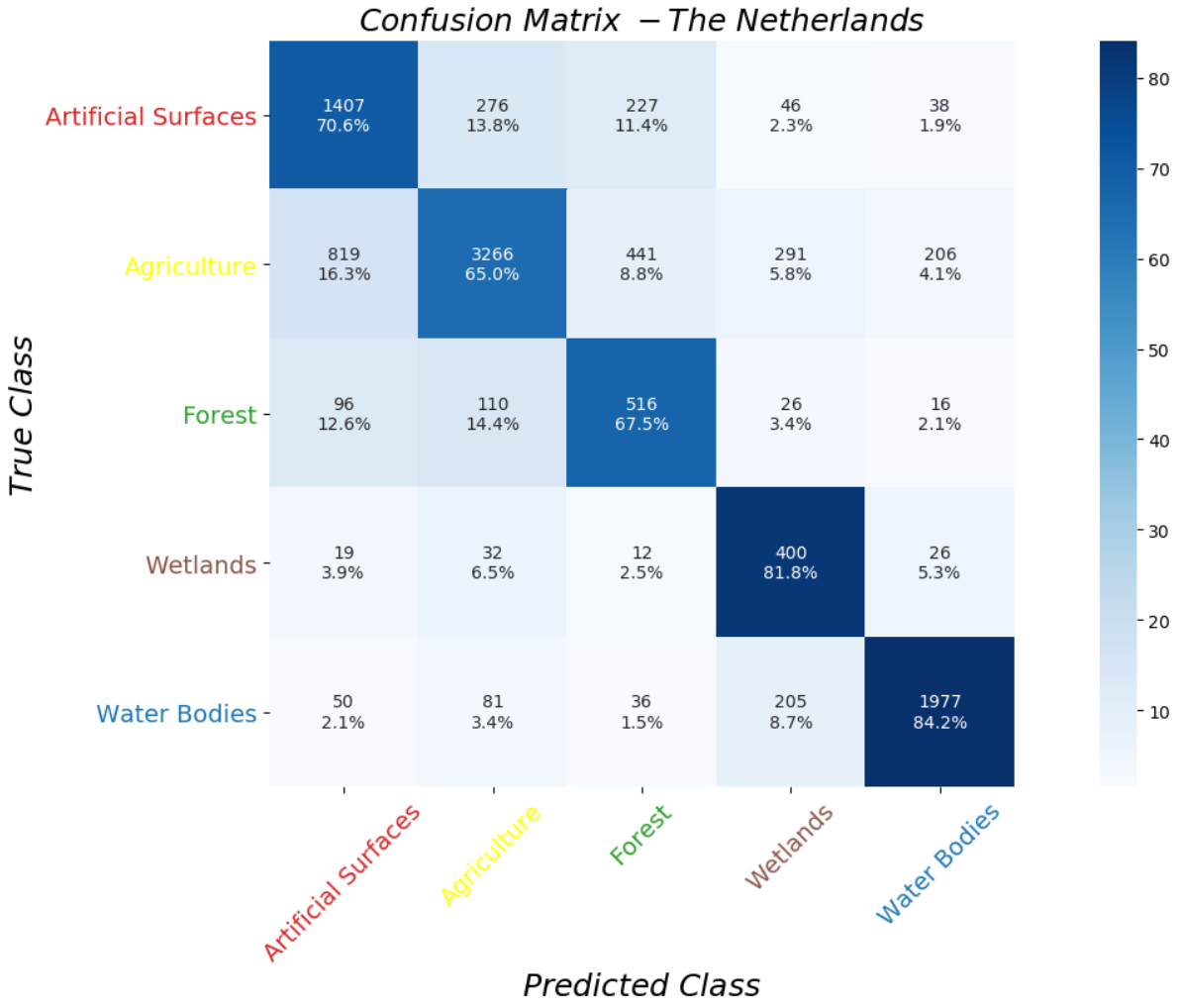


Figure 4.4: Classification Result - The Netherlands/ Belgium - Confusion Matrix

Artificial surfaces were classified with 70.6% accuracy. Considering the high variability of surface types within this class, from Dump sites to Sports and leisure facilities, this is considered a positive result. Artificial surfaces shares samples that are similar to other classes, both in surface type and the nature of the characteristic profile. In the case of the former, Artificial surfaces such as golf courses and sports fields are similar to a typical Agriculture profile. In addition, Artificial surfaces are often found in the Agriculture validation samples. For example, small roads or houses and agriculture buildings attributed to a farm. The highest classification error of 30.1% was indeed observed between these two classes (13.8% for Artificial and 16.3% for Agriculture). In fact, Agriculture was the class with the lowest classification accuracy at 65.0%. In addition to the discussed overlap with Artificial surfaces, Agricultural surfaces often find themselves on the fringes of other classes such as Forest and Semi-Natural Areas and Wetlands. With the Forest and Semi-Natural areas class, its useful to realise (referring to class surface breakdown in Figure 3.5), that in addition to Forests, this class also consists of natural grassland, transitional woodland shrub, beaches and sparsely vegetated areas, which all share similar properties to Agriculture. The classification error between Agriculture and Forest and Semi-natural areas was 8.8%. Its also necessary to mention that Agriculture had many more (5023) test samples than the other classes and is perhaps a more robust measure than, for example, Forest and Semi-natural areas (764), albeit at a lower accuracy.

Forest and Semi-natural areas was classified with an accuracy of 67.5%. Similarly to Artificial surfaces, Forest and Semi-natural areas has a large variability in surface types. From dense forests to beaches and bare rock. There tends to be a problem with classification for this class when the classification cell does not have at least one tree. In this case, the algorithm would often classify as Agriculture. This happened in 14.4% of test samples for this class. There were a significant amount of training samples in which the profiles between Forest and Semi-natural areas and Artificial Surfaces were similar. For example, in a particularly dense forest at around the height of a house. As discussed in Section 3.5.2, the peak detection algorithm was implemented in order to reduce this confusion, however there is still confusion in 24.4% (11.4% for Artificial and 12.6% for Forest) of samples between these two classes.

4.2 Feature Importance

This Section aims to decipher which features were most important in distinguishing one class from the others. The calculation by which the feature importance is measured is discussed in Sections 2.3 and 3.6.3.

While the methods used to analyse feature importance indeed provide interesting and valuable insights, the nature of the Random Forest algorithm means that it is difficult to understand the exact effects of the features on the result, given the fact that the forest has 500 trees, each with a tree depth of 30. By using the One vs All classification strategy, the question that is addressed is "How does a feature contribute to distinguishing one class from the others?"

The selection of optimal features is a field of study in its own right, whether one is attempting to select optimal features for the best classification accuracy, or for computational efficiency. Of course, the approach taken depends on the specific problem and the desired result/application. One would usually expect the process of feature selection to be performed in the creation and tuning of the model itself. As this research serves as a baseline for multi land type classification of ATLAS data, the aim is not necessarily to optimise the features selected (i.e only use the best features), but to understand which do and do not effectively contribute towards the classification of the 5 land types investigated.

The Section is separated according to the land types investigated. As there were 51 features implemented in total, the discussion will often refer to groups of features, most notably the groups that were introduced in Section 3.5, that is "Altitude-derived", "Histogram-based" and "Eigenvalue-based". Firstly, a few general conclusions are provided followed by a per class discussion to highlight the features that are uniquely important to each class. In order to remain concise, the feature importance results have only been included for Artificial Surfaces. For the rest of the classes a discussion will be provided and the results can be found in Appendix B. The reader may notice that the values of the feature importance's are not discussed. This is due to the fact they indicate the relative feature importance for each One vs All model applied and therefore the values for each class investigated are difficult to compare. In other

words, the numeric values for the feature importance provided in Figure 4.5, which are derived from the model classifying Artificial surfaces against all other classes, cannot be compared to the other feature importance results provided in the Figures in Appendix B, which compare each class in turn against the remaining four classes. The sum of feature importance in each model is equal to 1.

There is some level of general agreement for the feature importance's between the classes. For example, as expected, some of the eigenvalue based features that didn't contain the second principle direction, such as *Scattering*, *Anisotropy* and *Change_of_curvature* were not important features for any of the classes. Conversely, *Linearity*, *Planarity* and *Eigentropy* were all notably important for the classes that had significant variation in height, that is, Artificial Surfaces, Agricultural Areas and Forest and Semi-natural Areas. For the water classes, none of the eigenvalue-based features were of significant importance.

The histogram based features were particularly important for the Artificial Surfaces class, where the high strength of the second peak indicates artificial scattering surfaces (where the strongest/most persistent peak represents the underlying surface). Apart from this case, generally, the heights at which the peaks occurred (i.e *h_strong*, *h_second*, *h_peak_mean*) and the differences in height between between them (i.e *h_strong_second*), carried more importance as oppose to their persistence or difference in persistence (*dPers*). There were a few features that simply performed poorly and were ranked at or near the bottom of feature importance for all of the classes. These include the persistence of the weakest peak, *Pers_weak*, *Pers_weak_rel*, the features related to empty bins, *empty_bins* and *consec_max_bin_count* and the number of peaks, *num_peaks*.

The altitude based features were the most important feature set. Of course, different percentiles were more indicative for different classes, but in general *perc_95_5*, *perc_75_25*, *ninety_fifth_per* and *h_max* were always important features for all classes. In general, the lower percentiles, such as *fifth_perc*, *ten_perc* and *quart_perc* were less important than the higher percentiles. This is thought to be due to the fact that the differences in height and characteristic profiles between the classes is less pronounced closer to the surface, especially in The Netherlands, where the topography is relatively flat over most of the country.

4.2.1 Artificial Surfaces

Figure 4.5 shows the feature importance rankings for Artificial surfaces. It is important to note that the rankings are relative rankings and don't provide information regarding the quality of the model itself. The red bars provide the mean importance calculated from the sum of the gini-impurity for each feature over the forest. The blue lines indicate the standard deviation. A negative value indicates that a feature negatively contributed toward the classification. This can be seen in the standard deviation for some features. This can happen for a variety of reasons. It could be that the it was a poor test sample, for example, the surface that was seen by the laser footprints was not what the validation data indicated. An additional reason for negative feature importance could lie in the fact that there can often be some ambiguity in the distinction between classes. For example, a semi-natural grassland, where features from this surface would strongly indicate an Agricultural surface. In doing so, negatively effecting the classification. In this sense, features with a negative importance can also indicate that this feature can be one who's importance is shared between multiple classes.

The most important feature for Artificial surfaces is the relative strength of the second strongest histogram peak, *Pers_sec_rel*. In addition, the vertical distance between the two most defined scattering surfaces (strongest and second strongest peaks), *h_max_sec*, is within the 15 most important features. In general, the ground/underlying surface of a classification cell would produce the strongest peak. Artificial surfaces, in comparison to other classes, tend to have a relatively strong second peak due to the strong scattering surfaces found within their classification cells, such as roofs. These two features are only also of notable importance in the Forest and Semi-Natural class, understandable given the fact that it is the only other class with a high photon height range, leading to strong scattering surfaces at significantly different ranges.

Artificial Surfaces tend to have the biggest range of heights within the cells as well as the highest absolute

heights. This seems to be a discriminative characteristic as features such as *perc_95_5*, *h_var*, *h_max*, *h_std*, *perc_75_25*, *ninety_fifth_perc* and *h_range* all appear within the top 15 most important features. This is in comparison to Agricultural areas for example, where those altitude derived features that define the extremes in height are much less important.

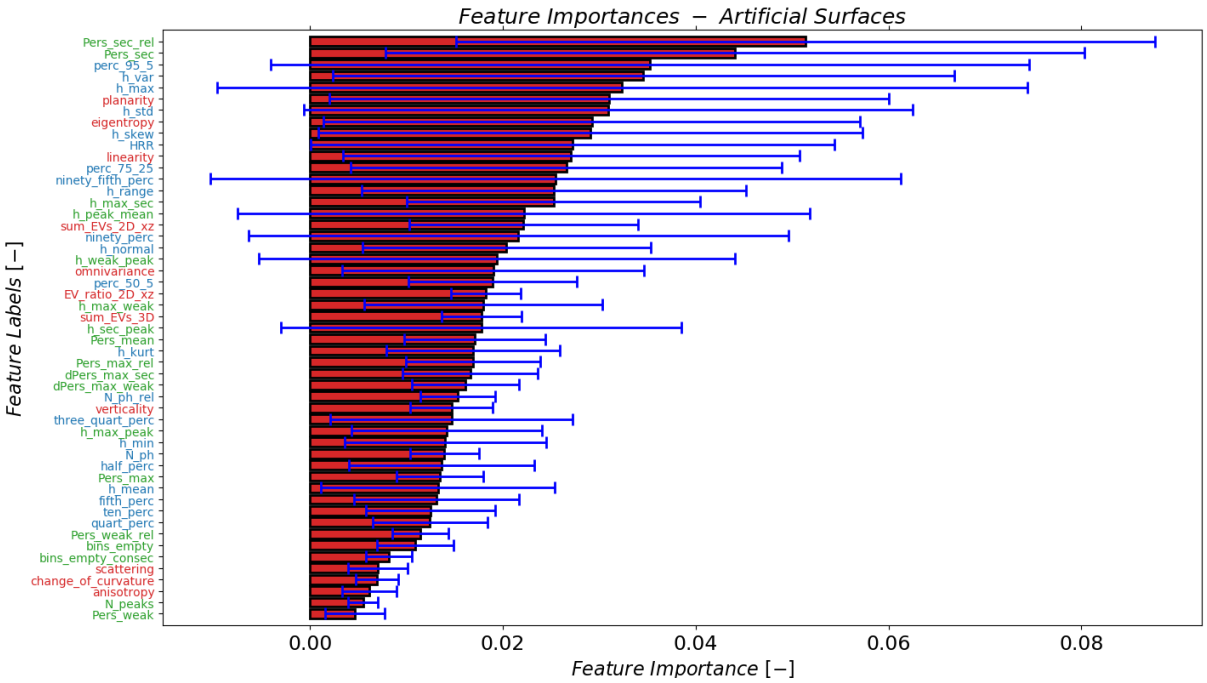


Figure 4.5: Feature Importance - One vs All Classification Strategy - Artificial Surfaces
51 features: 22 Altitude-derived, 18 Histogram-based, 11 Eigenvalue-based

With regards to the eigenvalue based features, "*planarity*", "*eigentropy*" and "*linearity*" are significantly important. As discussed in Section 3.5.3, it was expected that the eigenvalue based features that pertain to the second principle direction would be most important and this is indeed the case. Linearity and Planarity measure the difference in magnitudes between the first and second, and second and third principle directions respectively. It seems as if there is a distinguishable difference in feature values for these three eigenvalue based features as they are also significantly important for the agricultural areas class, whose characteristic profile is notably different.

4.2.2 Agricultural Areas

The importance of the features for Agricultural areas was significantly more evenly distributed than other classes, that is, the difference in importance between most of the features is less pronounced. As oppose to, for example, Artificial Surfaces and Wetlands which have a few features that are relatively strong and many that are much less so. Agricultural areas was the worst performing class and this could indicate that the algorithm is struggling to find features that can distinguish agricultural areas from the other classes. Another hypothesis is that as the agricultural class often contains a good deal of overlap with other classes (streams in fields, vegetation separating fields, farmhouses etc), meaning that a wide variety of features can aid in the distinction of this class, leading to a more evenly distributed feature importance.

However, there are still features that are uniquely important to this class. The $sum\_EVs\_2D\_xz$ is the most important feature for this class. It seems that the combination of the behaviour in the x and z plains is unique to Agriculture. Indeed, the characteristic agricultural profile of agriculture is unique in comparison to other classes. It is likely the random and unstructured returns from the low lying vegetation in this plain contribute towards $sum\_EVs\_2D\_xz$ being so important for this class. As a comparison, one would expect the variation in this plain to be structured and constant on an artificial surface and to a lesser extent, the same can be said for a dense forest canopy. Agricultural surfaces

provide uneven surfaces with varying vegetative densities and heights. This is thought to contribute toward the unique eigenvalues that constitute this feature.

h_normal is also uniquely important for this class. Likely due to its unique height range, if you consider that that this feature (and many other altitude derived features) will be similar for the pairs Artificial Surfaces and Forests, as well as Wetlands and Water Bodies. Agriculture ranges from just above the water level to within approximately $1m$ above that. Whilst Wetlands and Water Bodies are ranging from the water level to just above and of course, Artificial Surfaces and Forests are ranging to many metres above the surface.

The Agricultural Areas feature importance results can be found in Appendix B, Figure B.1.

4.2.3 Forest and Semi-natural areas

In comparison to Agricultural Areas, Forests have a less even distribution of feature importance over the 51 features, with the top 8 features sharing significant importance. The most important features for Forest and Semi-natural areas are a mix of the most important features from Artificial Surfaces and Agricultural Areas, that is, the features that indicate a larger range in height such as *h_std*, *h_var* and *perc_95_5*, as well as the most important feature from Agricultural areas, $sum\_EVs\_2D\_xz$. In the case of the latter, this could share the hypothesis of Agricultural areas that this feature represents the uneven ground associated with agriculture and Forest canopy floors.

While *h_std* is an important feature for Artificial Surfaces, it is uniquely the most important feature for Forest and Semi-natural areas. Conversely, the strength of the second peaks, $Pers\_sec\_rel$, is the most important feature for Artificial Surfaces and considerably less so for Forest and semi-natural areas. As vegetation (and uneven ground) is less reflective than flat and constant artificial surfaces, the peaks are less prominent, meaning that a more general statistic, such as *h_std* and *h_range* describing the entire cell profile contains more discriminative information for the Forest and Semi-natural areas class than the histogram peaks.

The Forest and Semi-natural feature importance results can be found in Appendix B, Figure B.2.

4.2.4 Wetlands and Water Bodies

It is appropriate to discuss Wetlands and Water Bodies together, as they share similarly important features, while the differences in feature importance are also telling of the differences in the surfaces covered by these two classes.

The most important features are, for both Water Bodies and Wetlands, dominated by altitude derived features such as $ninety\_fifth\_perc$, *ninety_perc*, $half\_perc$ and *three_quar_perc*. This makes sense as these percentiles will generally all be in a uniquely low range with respect to the other classes. The range in height is relatively much smaller than other classes as well as the average.

The algorithm actually distinguished between Wetlands and Water Bodies very well, considering how similar the profiles can be. Only 5.3% of Wetlands samples were misclassified as Water Bodies and 8.7% of Water Bodies were misclassified as Wetlands. Probably, this is attributed to the fact that, in general, the Wetlands class will sit slightly higher up in altitude than Water Bodies, which are more or less all at sea level. In addition to this, the often relatively uneven and vegetative surface of Wetlands compared to Water Bodies could change the form of the height profile to enough of an extent to allow for the successful discrimination between the two classes.

The only eigenvalue based feature of significant importance is the $sum\_EVs\_2D\_xz$ for the Wetlands class. Again, similarly to Agriculture, probably due to the uneven surfaces that are associated with the Wetlands class. In general, the eigenvalue-based features are not important for the distinction of water classes, likely due to the relatively low variance of the profiles with respect to the other classes. The feature importance results can be found in Appendix B, Figures B.3 and B.4 for Wetlands and Water Bodies respectively.

4.3 Misclassification Analysis

In order to better understand the capability of classifying land types with ATLAS data, as well as assessing the suitability of CORINE for validation, a detailed misclassification analysis has been performed. Out of the 10,619 test samples, 3053 (28.8%) were incorrectly classified. 1052 of these incorrectly classified samples (34.5%) were inspected in an attempt to determine the nature of the misclassifications. There were many different forms of error, however, these have been grouped into 4 high level error types for conciseness. Section 4.3.1 quantifies and provides a summary of the error types and sections 4.3.2 to 4.3.5 provide definitions of the error types as well as provide some real examples.

Four high level error types were identified in the analysis of the classification cells that predicted a different class than that of CORINE. Each error type discussed can be broken down into lower levels of error types, however the presentation of them in this report was considered redundant in terms of the drawn conclusions. The error types were identified by superimposing the classification cells onto a map, where the predicted and true classes were indicated. By visually analysing the surface content of the classification cells relative to where in the cell the footprint of the laser lay, it was possible to determine whether or not the misclassification was actually a classification error or not. Its necessary to mention that the surface images that were used for analysis were taken at various points throughout 2019, whereas the validation data was collected in 2018. In addition, the weather condition above each classification cell at the moment of acquisition is unknown. As previously discussed, the omission of classification cells that didn't meet the threshold for the minimum number of photons is considered to sufficiently remove the effect of weather for the purposes of this research. The determination and assessment of the algorithmic misclassifications discussed in this section have been performed by a human, and are therefore open to a certain degree of subjectivity.

It is first necessary to briefly define the error types, a more detailed description (as well as visual examples) of each error type can be found in Sections 4.3.2 to 4.3.5:

Type A	A genuine predictive confusion of the classification model
Type B	The laser footprint passes over multiple classes within the classification cell and the classification model predicts the minority class seen by the laser
Type C	The laser footprint passes over multiple classes within the classification cell and the classification model predicts the majority class seen by the laser
Type D	The validation data is considered incorrect

4.3.1 Error Analysis/ Overview

Of the 10619 test samples, 3053 were incorrectly classified. Just over a third, 1052 samples, were visually checked in order to quantify the types of error that were occurring. The results of this sub-sample analysis can be found in Table 4.1.

	Artificial Surfaces	**Agriculture**	**Forest**	**Wetlands**	**Water Bodies**	**Total**	**Total [%]**
Type A	169	313	70	33	101	**686**	**65.2**
Type B	19	96	4	0	7	**126**	**12.0**
Type C	28	151	13	1	8	**201**	**19.1**
Type D	7	19	12	0	1	**39**	**3.7**
Total	**223**	**579**	**99**	**34**	**117**	**1052**	

Table 4.1: Misclassifications categorised by Error Type

The rows represent the error types described from Sections 4.3.2 to 4.3.5 and the columns represent the land types. The total errors can be seen in the last two columns of Table 4.1. Only 65% of the misclassifications analysed were deemed to be genuine confusions (Type A). 3.7% of the validation error

were considered to be erroneous, that is, the validation data was considered to be incorrect. 31.1% of the misclassifications consisted of errors arising from the laser footprints passing over multiple classes within the classification cell. This was separated into Type B error for the prediction of the minority class seen by the laser, and Type C error for the prediction of the majority class seen by the laser, where Type B and Type C errors amounted to 12.0% and 19.1% respectively.

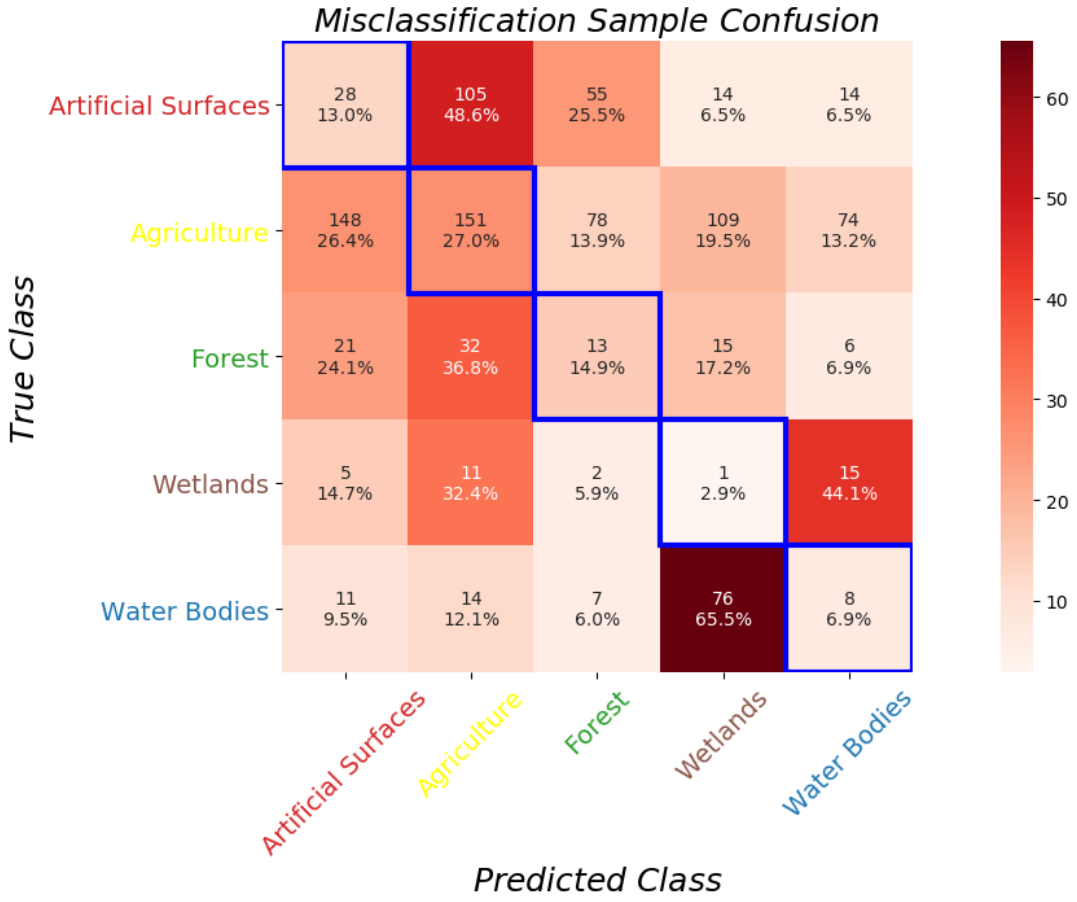


Figure 4.6: Misclassification Confusion Matrix - 1052 misclassified samples.
Type C errors, which are now considered correct classifications, populate the diagonals of the confusions matrix, represented by the blue border. As this is a confusion matrix of the error, the diagonals would usually remain unpopulated. Therefore, this confusion matrix can be considered as indicating the reduction of error due to the visual inspection of the incorrectly classified cells.

In order to re-asses the total accuracy of the model, Figure 4.6 provides the confusion matrix of the misclassified cells. While one could consider Type B errors to be partly correct, for this analysis, they are still considered misclassified as they were unable to predict the majority class. Type D errors have been removed. Type C errors, the predictions of which are now considered correct after visual inspection, can be seen along the diagonal of the confusion matrix, indicated by the cells with the blue border. Of course, if one were to plot the confusion matrix of all the misclassifications without this analysis, the diagonal cells would be unpopulated. This confusion matrix therefore provides a more comprehensive view of the error, where the diagonals provide an indication of the per class added accuracy that could be expected from the final result if all the errors were to be visually analysed.

Overall, with a 19.1% Type C error, one could expect the total accuracy of the model to be increased by 5.5%. The largest source of accuracy gain, or Type C error, came from the Agriculture class. This is due to the frequent in cell overlap this land type shares with the other classes. Often, a large part of the

cell would consist of an artificial surface such as a greenhouse, or the edge of a forest and the algorithm would predict accordingly.

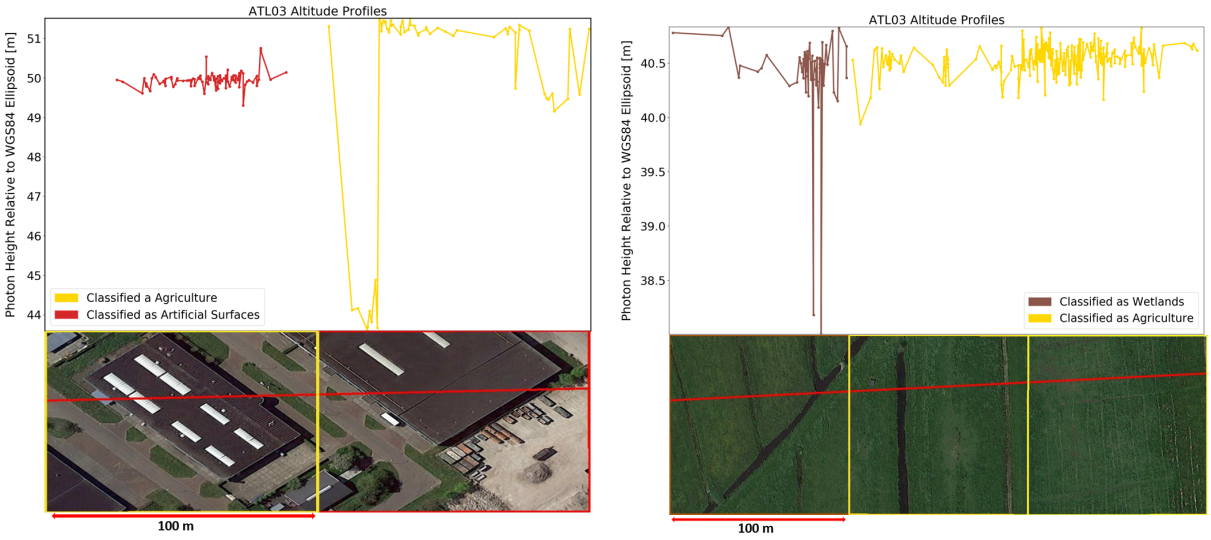
It is expected that Type B error, of 12.0%, could largely be reduced by either classifying at a higher resolution (smaller classification cells) or using smaller segments of the classification cells to vote for the overall outcome of that cell. In this way, the minority class that is predicted would not carry as much weight in the classification, as the effects of its feature values on the total results of the cell would be reduced. Imagine that of a $100m$ classification cell, $90m$ consists of a typical agricultural field and the final $10m$ an industrial greenhouse. The industrial greenhouse will largely alter the feature values for that cell, increasing the average and altering the percentile heights as well as adding defined scattering surfaces (peaks) in the histograms, leading to an incorrect prediction.

As with Type B error, Type C error is largely a results of the measurement technique and validation data. These errors can only be avoided by visual inspection and would require a higher resolution data set to be removed automatically.

4.3.2 Type A Error - Genuine Predictive Confusion

A Type A error is defined as a genuine confusion, that is, the laser passes mostly over a cell of which the validation is considered correct, yet classifies the cell as a different class. Of course, these errors give the most insight into the weaknesses of the classification algorithm itself as oppose to the flaws with validation data and its attributes. Therefore, a more detailed analysis will be provided of this error type.

Of course, there are many different types of genuine confusion, however, only some of the most common errors for each class will be presented, with the intention to remain concise, while providing the most significant trends of confusion. Figures 4.7 and 4.8 provides some individual cases of confusion. In each sub-figure, profiles are provided, with the classifications cells and laser footprints below. The profiles and respective cells are coloured to match the predicted class.



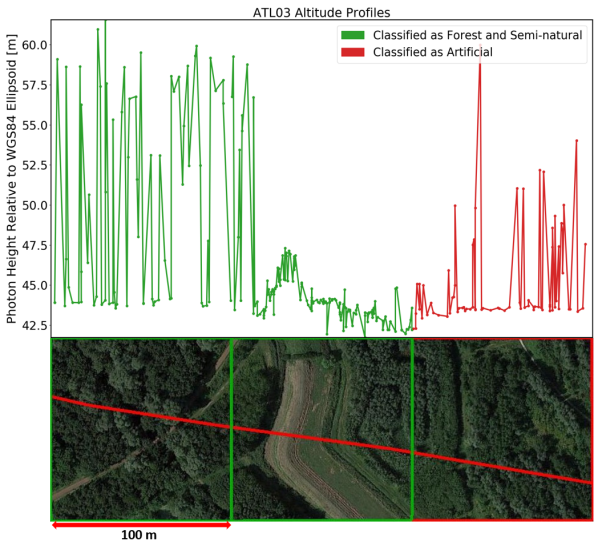
(a) True: Artificial Surfaces
Predict: Agricultural Areas

(b) True: Agricultural Areas
Predict: Wetlands

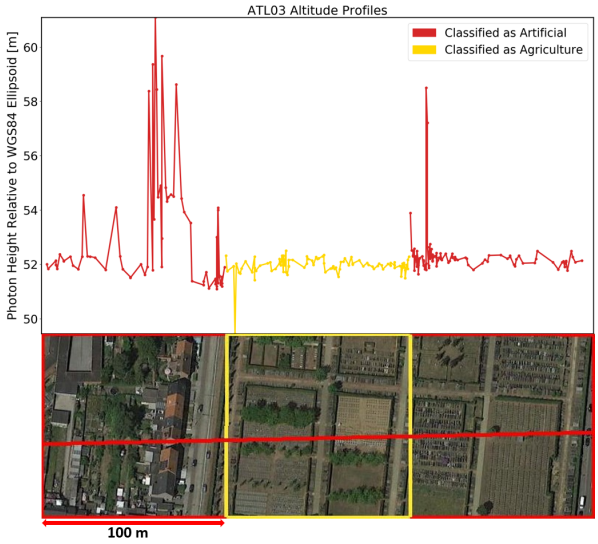
Figure 4.7: Type A Misclassification examples.
Photon height profiles of $100m$ *classification cells, with the location and footprint of the cell below. The intersection of a vertical line drawn from any point of the profile to the footprint is the location at which a photon was recorded.*
Key: **Artificial Surfaces**, **Agriculture**, **Wetlands**

In Figure 4.7a, the left cell is misclassified as agriculture. Looking at the profile of the left classification

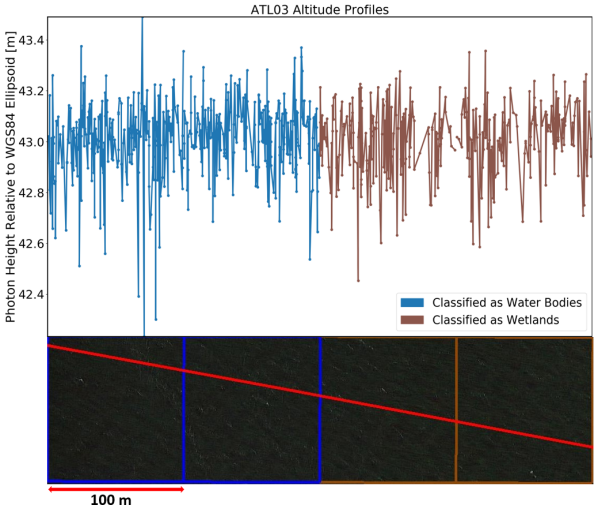
cell in Figure 4.7a, it seems that photons were only returned from the warehouse, and not the underlying surface on either side. Clearly, the roof wasn't flat and a profile similar to a characteristic agriculture profile was returned. Whereas, in the cell on the right, photons were also returned from the ground, providing two scattering surfaces $7m$ apart. While one may expect that the high average height of the left cell relative to most agricultural surfaces would be an indicator to not classify as such, it seems that in this case, an additional scattering surface was required. In this case, the conditions were clearly not optimal as only part of the cells profile was returned. Of course, the higher the photon return within a cell, the more information the algorithm has for classification.



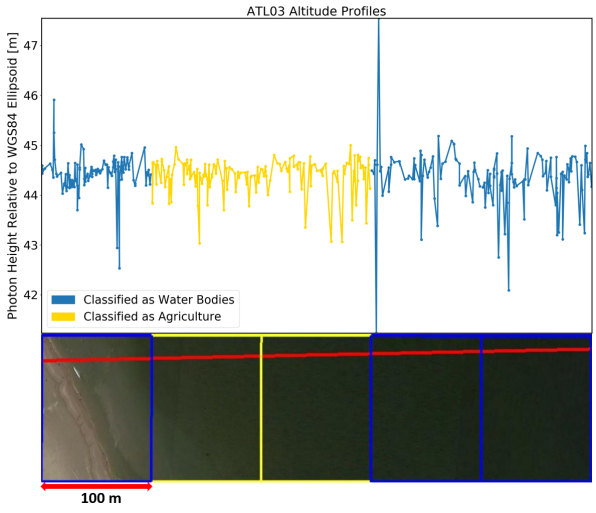
(a) True: Forest and semi-natural
Predict: Artificial Surfaces



(b) True: Artificial Surfaces
Predict: Agriculture



(c) True: Water Bodies
Predict: Wetlands



(d) True: Water Bodies
Predict: Agriculture

Figure 4.8: Type A Misclassification examples.
Photon height profiles of example $100m$ *classification cells, with the location and footprint of the cell below. The intersection of a vertical line drawn from any point of the profile to the footprint is the location at which a photon was recorded.*
Key: **Artificial Surfaces**, **Agriculture**, **Forest**, **Wetlands**, **Water Bodies**

Figure 4.7b shows a common error of "class mixing". In this case, the left classification cell is incorrectly

classified as Wetlands. Probably due to the fact that the water body at the end of the cell was particularly reflective, introducing properties attributed to Water Bodies/Wetlands into the profile, such as the downward pointing spikes. In the next cell, there is also a small stream, however for some reason is not as pronounced as in the previous cell. Class mixing was a common reason for confusion. Some other examples of note were, sports fields and golf courses classified as Agriculture or Forest and semi-natural, and urban areas with a high density of trees classified as Forest and Semi-natural.

Figure 4.8a show a Forest and Semi-natural area cell misclassified as Artificial Surfaces (third from the left classification cell). While there is clearly an artificial surface in the middle cell, it is correctly classified as Forest and Semi-natural. It is thought that the relatively flat surface in the incorrectly classified cell has contributed toward the Artificial Surfaces classification. Whereas in the other two cells, the underlying surface is bumpy or ascending/descending, indicating a natural surface. As previously discussed, the profiles of Artificial and Forest and Semi Natural can often be similar and there was a 24% error rate between these two classes.

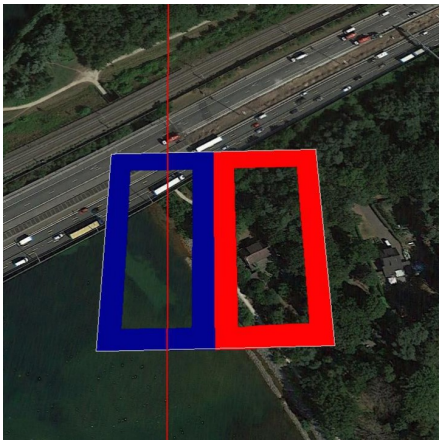
Figure 4.8b provides another example of class-mixing. In the middle cell, the footprint passes over a part of a cemetery without raised objects/structures, but does pass over vegetation, resembling an Agricultural surface. In the cell to the right, also a cemetery, it seems that a tree or other structure was recorded, indicating a significant height difference and therefore an Artificial Surface.

The nature of errors of the final two sub-figures, that is, Figures 4.8c and 4.8d, are somewhat more ambiguous. In both cases, Water Bodies has been misclassified. It is not immediately obvious from the profiles why this was the case. In Figure 4.8c, the final two cells are misclassified as Wetlands and in Figure 4.8d two of the middle cells are misclassified as Agriculture. It is often the case for the water classes that photons are returned from somewhere below the water surface and in shallower waters the photons can be returned from the ground surface. This could be the case in Figure 4.8d, as the data is clearly taken in shallow waters, then again, the cells either side are correctly classified as Water Bodies. In Figure 4.8c the only noticeable difference is the more staccato nature in the along track direction of the profiles classified as Wetlands, that is, there is less of a constant photo return in some places. This is often seen in Wetlands classes, as there are many Wetland samples that have vegetation within them, which can be much less reflective than water. Perhaps the break in reflectivity has changed eigenvalue based features enough to indicate Wetlands. The errors shown in Figures 4.8c and 4.8d show the drawback of the Random Forest algorithm as it is difficult to say for certain what is causing the misclassificaiton.

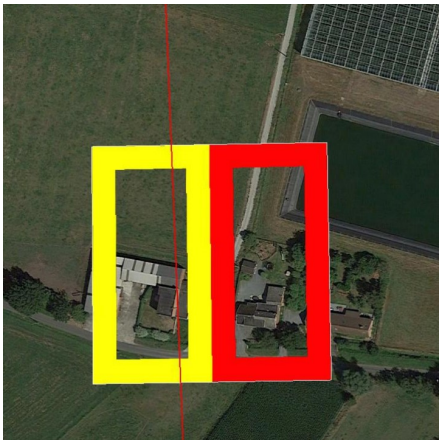
4.3.3 Type B Error - Minority Class Confusion

A Type B error is defined as one in which the land type designated by the validation set is considered correct, that is, 50% or more of the classification cell consists of the indicated class. However, the laser footprint passes over a different class for a minor proportion of the cell and is such classified as the minor class. The difference with class mixing, which was discussed for Type A errors, is that class mixing represents ambiguity between class definitions, such as a vegetative surface in an urban landscape (i.e garden). Whereas, in Type B errors, there are distinctly different classes appearing within the classification cell, such as the house and field in Figure 4.9c.

(a) Majority/ True: Artificial Surface
Minority/ Predict: Wetlands



(b) Majority/ True: Water Bodies
Minority/ Predict: Artificial Surfaces



(c) Majority/ True: Agriculture
Minority/ Predict: Artificial Surfaces

(d) Majority/ True: Agriculture
Minority/ Predict: Forest

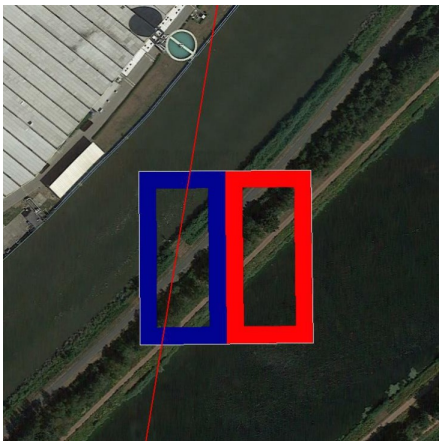
Figure 4.9: Type B Misclassification examples.
The left-hand side of the cell indicates the True class and the right-hand the predicted class. In the case of Type B errors, the minority class has been predicted, that is, the class which is seen the least by the footprint.
Key: **Artificial Surfaces**, **Agriculture**, **Forest**, **Wetlands**, **Water Bodies**

Figure 4.9 provides 4 instances of Type B error. The colour of the left hand side of the classification cell indicates the true class and the colour on the right hand side indicates the predicted class. In all 4 cases, it can be seen that the minority class has been predicted by the classification algorithm, that is, the class which is passed over the least by the laser footprint. For example, in Figure 4.9b, the laser track passes over a raised road for only a small part of the cell, however, the effect on the features due to this surface seem to be enough to predict Artificial Surfaces. This is also the case for the other examples

in Figure 4.9, that is, the minority class dominates the prediction results. In the case of Type B errors, one would expect an increase in performance if a higher resolution of classification cells were used, that is, for example, if $10m$ classification cells were used. In this case, one would expect the algorithm to identify the correct class either side of the in-cell class boundary between, for example, the bridge and road in Figure 4.9b.

4.3.4 Type C Error - Majority Class Confusion

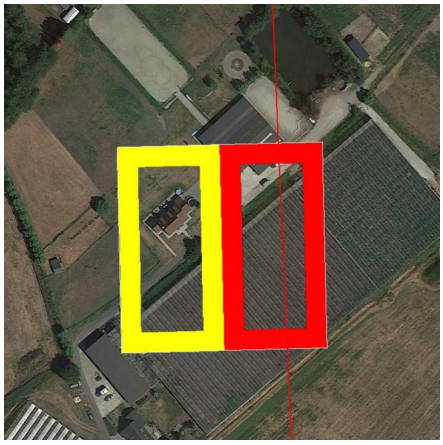
Type C errors are defined as those in which the laser footprint passes mostly (or fully) over a different class than the true/CORINE class. In this sense, one could not expect the algorithm to predict the correct class as it is not seen by the laser. A cell is considered a Type C error if, upon visual inspection, the prediction is considered correct, that is, it predicts the majority class seen by the laser. Type C error can for the most part be considered as the opposite of a Type A error, however, in some cases, such as Figure 4.10d, none of the cell contains the typical surface of the indicated validation class.



(a) Predicted: Artificial Surfaces
Validation Class: Water Bodies

(b) Predicted: Wetlands
Validation Class: Agriculture

(c) Predicted: Forest
Validation Class: Agriculture



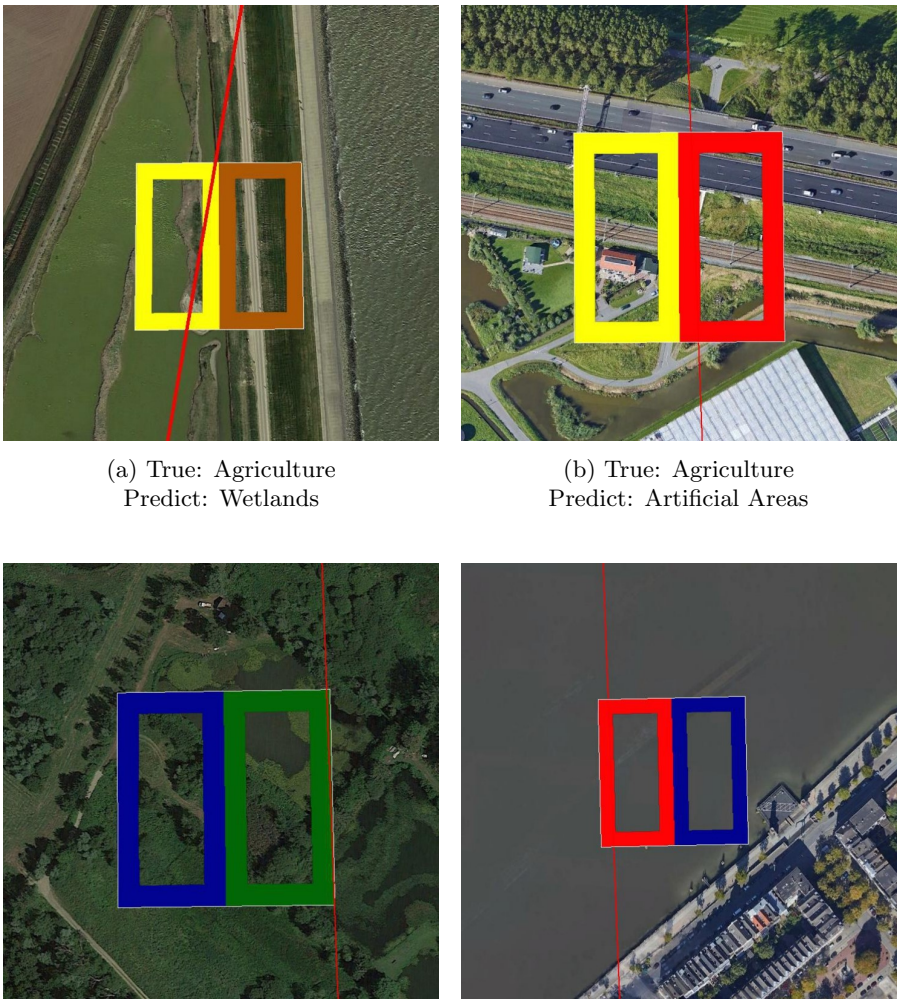
(d) Predicted: Artificial Surfaces
Validation Class: Agriculture

Figure 4.10: Type C Misclassification Examples.
The left-hand side of the cell indicates the True class and the right-hand the predicted class. In the case of Type C errors, the majority (or only) class has been predicted, that is, the class which is seen the the most/fully by the footprint.
Key: **Artificial Surfaces**, **Agriculture**, **Forest**, **Wetlands**, **Water Bodies**

Figure 4.10 provides 4 instances of Type C error. In all 4 cases, it can be seen that either none or most of the footprint within the classification cell passes over the true class. For example in Figures 4.10c, the footprint passes almost completely over a forested area, correctly classifying it as such. However, the true class, understandably, is Agricultural areas. Figure 4.10a is more similar to a Type A error, where there are two distinct classes that the footprint passes over, in this case, Artificial Surfaces and Water Bodies. In this case the cell is incorrectly classified as Artificial surfaces. However, upon visual inspection, this results is considered correct and it is therefore designated as a Type C error. Type C errors, as discussed in the overview in Section 4.3.1, are considered in this misclassification analysis to be correctly classified. In the case of Type C error, it is not necessarily that the validation data is incorrect, but that it is not fully compatible with the laser footprints.

4.3.5 Type D Error - Incorrect Validation Data

A Type D error is one in which the validation class is simply incorrect. That is, none of the cell constitutes a surface associated with the indicated class. As discussed, the validation data was collected in 2018, whereas the ATLAS and map images used to check the results were acquired at different points in 2019. It could be that the land type has changed in between theses two acquisition dates or it could be that the validation was simply incorrect.



(a) True: Agriculture
Predict: Wetlands

(b) True: Agriculture
Predict: Artificial Areas

(c) True: Water Bodies
Predict: Forest

(d) True: Artificial Surfaces
Predict: Water Bodies

Figure 4.11: Type D Misclassification examples.
The left-hand side of the cell indicates the True class and the right-hand the predicted class. In the case of Type D errors, the indicated True class is considered to be incorrect.
Key: **Artificial Surfaces**, **Agriculture**, **Forest**, **Wetlands**, **Water Bodies**

Figure 4.11 shows four instances of Type D error. In all cases, it is obvious that the validation class (left hand side of the classification cell) is incorrect. In Figures 4.11b and 4.11d, the classification cell consists fully of the predicted class (right hand side of the classification cell).

5. Conclusion & Recommendations

Section 5.1 presents the conclusions to the research questions set out in Section 1. Recommendations for further research and aspects that were either out of the scope of this research or were not able to be implemented due to time constraints are presented in Section 5.2.

5.1 Conclusion

The wider goal of the research presented in this report was to better understand the behaviour of ICESat-2 geo-referenced photon data. The observation and measurement of individual photons on the scale provided by the ATLAS instrument is a novel application in satellite laser altimetry and the motivation behind the study was to acquire knowledge of how ICESat-2 interacts and behaves when illuminating different land types. The ability to measure the vertical structure of the footprint offers a contrasting and unique perspective in comparison to the traditional use of spectral information to classify land types.

Referring to the first research question, namely, "How effectively can ICESat-2's ATL03 global geolocated photon data be used to classify between the land types Artificial Surfaces, Agricultural Areas, Forest and Semi-natural area, Wetlands and Water Bodies using the random forest machine learning technique?". The test considered all available ICESat-2 data from the fifth repeat orbit, from September 26^{th} 2019 to December 26^{th} 2019. The 101,681 applicable samples that were obtained for this period were separated into test and train data, of which 10619 were used for the testing. The overall accuracy of the classification model was 71.2%. This overall result considers the 5 land types that were classified: "Artificial Surfaces", "Agricultural Areas", "Forest and Semi-natural Areas", "Wetlands" and "Water Bodies". The water classes were the most successful land type, with Wetlands and Water Bodies achieving an accuracy of 81.8% and 84.2% respectively. Agricultural Areas was the worst performing class at 65.0%. Artificial Surfaces and Forest and Semi-natural areas achieved accuracies of 70.6% and 67.5% respectively. A large source of misclassification was found to be due to the validation data and the discretization size of the footprints into classification samples. After a visual analysis was performed on the misclassification samples, it was found that only 65% of the total errors were deemed to be genuine confusions. By visually expecting the misclassification samples, it was found that the accuracy was a minimum of 5.5% greater, leading to a total accuracy of 76.7%. In addition, it is expected that if the size of the resolution of the classification cells were reduced (i.e classify $10x10m$ cells), one could expect an additional gain in accuracy of 3.5%, resulting from the ability of ATLAS's lasers to better detect class boundaries.

The use of the random forest model has allowed the discovery of which features are important toward the classification of the 5 land types investigated. Unlike other classification models such as neural networks, the features are chosen by hand and this greatly contributed toward the understanding of the signal. Saying that, a detailed analysis of the exact behaviour of the forest was difficult due to the 500 trees included.

The second research question aimed to decipher "Which features unique to ATL03 geolocated photon data are best at distinguishing land types from one another?". Three different feature sets were used for the classification, namely, "Altitude-derived features", "Histogram-based" and "Eigenvalue based features". The importance of a feature was measured in the features ability to discriminate between classes, measured using the summation of the gini-impurity values over the classification forest. In general, the inclusion of the three feature sets was validated. There were features from all three features sets that were important for each class investigated. In general, the altitude derived features were highly ranked for all classes. Statistical features that described the distribution toward the higher end of the profiles were the most important, that is, the 90^{th} and 95^{th} percentile, the difference between the 95^{th} and 5^{th} percentiles and the maximum height of the photons. While the altitude derived features were successful in discriminating between all classes, there were certain features that were of particular importance to certain classes. The eigenvalue based features "Linearity", "Planarity" and "Eigentropy", derived from

the principle directions of the ATLAS profiles, were of significant important for the classes that had the most vertical variation their profiles, that is, Artificial Surfaces, Agricultural Areas and Forest and Semi-natural Areas. Eigenvalue-based features were not significantly important for the Water classes, whose profiles were generally represented by a single flat surface. The histogram derived features were particularly successful at identifying strong/persistent scattering surfaces. The definition of the second strongest scattering surface, that is, the strongest surface after the underlying ground surface, was the most important feature for Artificial Surfaces, while the average height of the scattering surfaces was the best feature overall from this feature set. There were a collection of features that were relatively much less important for all of the classes. Most notably, the eigenvalue-based features "change_of_curvature", "scattering" and "anisotropy" and the histogram features that described properties of the weakest/least persistent peak as well as those counting the empty histogram bins of a profile. All of the altitude derived features were of significant importance for at least one of the classes.

The third and final research question was, "What technique can be used to validate the classification?". The CORINE validation set was not an ideal dataset for the validation. While it's use in this research has provided satisfactory answers to the research questions proposed, there are also some aspects that have negatively effected the results. As discussed, upon visual inspection, it was found that the classification algorithm had actually performed better than the validation set indicated. The extent of this was quantified in the analysis of the misclassifications, where four different high level error types were discussed. Three of these error types were due to the incompatibility of the validation with the ATLAS footprints. ATLAS footprints will always require a custom validation strategy as the footprints are essentially continuous slices of the Earths topography, that is, they collect data only at the locations of the laser footprints. As the CORINE classification cells are 100x100m, it was often the case that the footprints passed over a different surface within the classification cell than indicated by the validation data. After a visual inspection, it was found that that only 65.2% of the misclassification errors were genuine confusions, the other 34.8% resulting from the discussed problems with the validation data. While the effect on the accuracy could be quantified with visual inspection of the misclassifications, the overall effect this had on the training of the model is unknown. CORINE was appropriate for showing that land types could successfully be classified with the use of ATLAS data, however, it is clear that superior results could be obtained with the use of a higher resolution and land cover validation dataset.

To conclude, this is the first time that ATLAS footprint have been used to classify multiple land classes. A classification strategy was proposed that segmented the continuous ATLAS footprints into 100m classification cells that were aligned with the CORINE 2018 land cover map. An acceptable accuracy of 71.2% was achieved, however, it was shown after an analysis of the errors that the true accuracy was 76.7%. Three features sets were used, namely, "Altitude-derived", "Histogram-based" and "Eigenvalue-based". While the simple statistics of the profiles defined within the altitude derived features were in general most important, there were features from histogram-based and eigenvalue based that were consistently important for discriminating between all classes. There is strong evidence that ATLAS data could be used to successfully classify at a higher resolution.

This result certainly provides motivation for further research of ATLAS data for land type classification. It is clear that discriminative information can be extracted from ATLAS profiles and this can add valuable input in combination with traditional spectral land classification methods. For example, it is thought that the ATLAS profiles have the potential to detect land classes at a higher resolution. This would be a particularly useful tool in the discrimination of vegetation types, such as grassland and forest. This functionality can also benefit many wider scientific users of ATLAS data. The detection of the class boundaries (and land type either side of them) in a variety of landscapes of interest is of worth and unique to ICESat-2 data, which is the only non-commercial satellite to offer repeated global measurements with a sub meter resolution. For example, the detection of the exact boundaries of an alpine lake will be beneficial to both the mapping of the local environment as well as the study of the lake itself. In the cryospheric environment, that which defines the predominant area of study of ICESat-2, the detection and definition of land type boundaries between rock, ice and snow are key to maintaining a continuous record of the short and long term changes that are occurring there.

5.2 Recommendations

From the research presented in this report, it is clear that ATLAS geo-referenced photon data can be used to classify land types. As this was a baseline classification, that is, it is the first time multi-class classification has been performed with ATLAS data, there are some areas in which it is thought that the classification could be improved. This Section will propose some recommendations that are thought to be useful areas of further research for ATL03 data.

In terms of features, there were many ideas that were not implemented. Firstly, only photons with a "high" signal confidence were used for the classification. However, there are many more photons available that are classified as "noise", "low" and "medium". It is unknown whether the medium or low confidence photons would strengthen the signal or negatively effect the results. [Zhang et al., 2020] performed a binary classification on Snow and Rock by only using noise photons, implying that there is valuable information even in the lowest confidence of photons. It is thought that more complex features could be derived from the signal, such as the co-efficients of various degrees of polynomial used to model the profiles. Of course, the inclusion of more complex features will provide a trade off with computation time.

As discussed, the aim of this report was not to optimise the classification results. However, it would certainly be interesting to do so. It is clear from the feature importance results that there are features that do not contribute significantly toward the final results. An optimal selection of the features presented in this report could improve the results, or the removal of less important features would create computational resources to test additional features or extensions of current features. For example, in this research, a histogram bin size of 20 is used, however, it is thought that additional information could be extracted with the use of various different bin sizes at the same time and/or a custom bin size depending on the range of height within the cell.

In addition to the feature selection, there were aspects of the classification itself that could also be optimised. For example, it was shown that the classification model was sensitive to the sample distribution used in the training. It was chosen to use an equal amount of samples per class, in order to create a balanced training, however there is clearly an optimal sample distribution for this classification that may not have been found. An optimal selection of the sample distribution or use of sample weighting would likely improve the results, especially as much of the training data was thrown away as the amount of samples per class used in training only matched the amount of the class with the lowest amount. In addition to the sample distribution, the hyper-parameters of the classification model were not rigorously tested, mainly due to the fact that there was not a significant effect on the model performance of the parameters that were tested.

With respect to the classification strategy, it is evident that ATLAS data will perform better when used to classify areas at a higher resolution than done so in this research. In fact, the use of $100m$ classification cells actually negatively impacted the results, that is, it was often the case (Type B and C error), that there were different surfaces within the classification cell. When a surface such as a house or tree is appearing in an otherwise flat landscape, even if only for a small portion of the cell, it has a large effect of the feature values and the classification algorithm often classified the cell incorrectly. However, due to the wider success of the algorithm, it is thought that the algorithm could successfully distinguish between a change in surface type. Classifying grid cells may not necessarily be the best way to classify land with ATLAS data, in fact, looking at the behaviour of the profiles, it is thought that it would be possible to actually find the boundaries between surfaces, indicated by, for example, observing the moving change in feature values across the landscape.

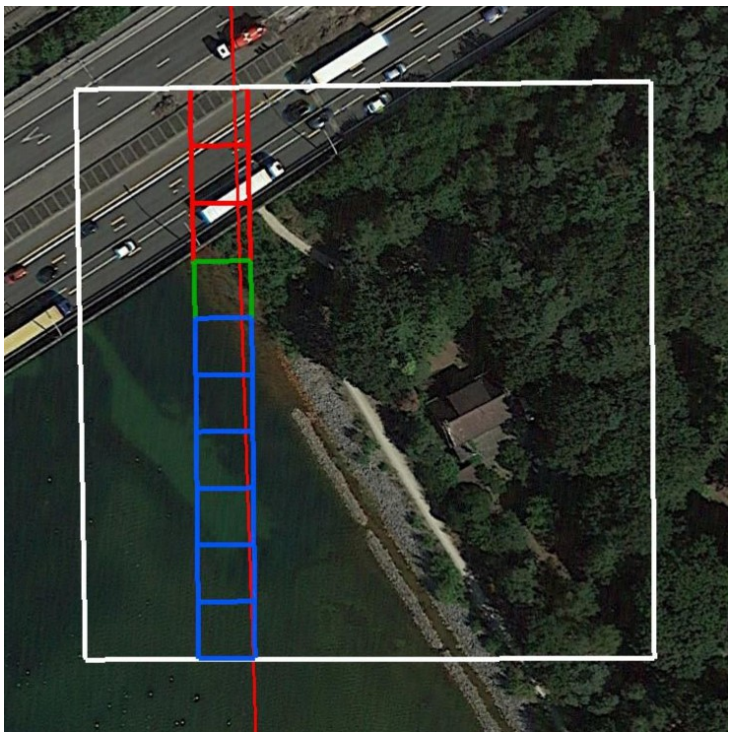


Figure 5.1: Example of hypothesised potential of the classification of higher resolution ($10x10m$) grid cells. CORINE class = Water Bodies, Predicted class: Artificial Surfaces
The coloured cells indicated what would be the correct classification of the grid cells. The overall prediction for the $100m$ *CORINE grid cell would be a vote of the smaller grid cells, which in this case would give the correct classification of Water Bodies*
Key: **Artificial Surfaces**, **Forest**, **Water Bodies**

There are two options with which this problem could be addressed. Firstly, would be the use of the same validation data, but classify at a higher resolution. Figure 5.1 provides a visual example of this, where $10x10m$ grid cells have been superimposed within the $100x100m$ CORINE grid cell. This example was first presented in 4.3.3 as a Type B error. The effect of the bridge at the end of the grid cell changed the features of the classification cell to enough of an extent to classify the cell incorrectly as an artificial surface. If one were to classify smaller grid cells, where each $10m$ cell would vote on the prediction of the $100m$ cell, this effect would be removed and the correct prediction would be presented. Of course, this assumes that the classification model would correctly classify the smaller cells.

A second method to combat the issues with the validation data would be to use a higher resolution validation dataset. Figure 5.2 provides an example of a validation data set that may be more appropriate for the classification of ATLAS data due to its higher resolution of surface classification. The Landelijk Grondgebruiksbestand Nederland (LGN2018) provides land type raster data at $5x5m$ pixel resolution. It would certainly be interesting to perform the same classification on classification cells such as this, given the expected increase in performance from the higher resolution grid cells outlined in this report. Unfortunately, the most recent version of the LGN is not open source and it's use would come at a cost.

Finally, it would be interesting to test this model on different land types than presented in this research. For example, the cryosphere is the main area of intended scientific study for ICESat-2 and the detection of its land types, such as ice, snow and rock could provide significant benefit for users interested in those regions. In addition to the region of study and land types under consideration, it is expected that the temporal acquisition of the data could have an effect on the results. The research presented here was taken from Autumn 2019, however it would be interesting to study the effect of seasons on the performance. For example, would the classification of forests perform better of worse when there are leaves on the tress?

Figure 5.2: Landelijk Grondgebruiksbestand Nederland (LGN2018), The Netherlands [Hazeu, 2019]. Each pixel is $5x5m$ and the different colours indicate a different land class.

Bibliography

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[Chawla et al., 2002] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

[Copernicus Land Monitoring Service, 2019a] Copernicus Land Monitoring Service (2019a). CLC 2018. Cited 07-08-2020 from: https://land.copernicus.eu/pan-european/corine-land-cover/clc2018?tab=metadata.

[Copernicus Land Monitoring Service, 2019b] Copernicus Land Monitoring Service (2019b). Corine Land Cover Nomenclature Guidelines. Cited 07-08-2020 from: https://land.copernicus.eu/user-corner/technical-library/corine-land-cover-nomenclature-guidelines/docs/pdf/CLC2018_Nomenclature_illustrated_guide_20190510.pdf.

[Criminisi et al., 2011] Criminisi, A., Shotton, J., and Konukoglu, E. (2011). Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning.

[Demantké et al., 2011] Demantké, J., Mallet, C., David, N., and Vallet, B. (2011). Dimensionality based scale selection in 3d lidar point clouds. In *Dimensionality based scale selection in 3D lidar point clouds*.

[Drummond et al., 2003] Drummond, C., Holte, R. C., et al. (2003). C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, volume 11, pages 1–8. Citeseer.

[Duong, 2010] Duong, H. (2010). *Processing and application of icesat large footprint full waveform laser range data*.

[epsg.io, 2019] epsg.io (2019). ETRS89-extended/ LAEA Europe. Cited 08-09-2020 from: https://epsg.io/3035.

[Feranec et al., 2016] Feranec, J., Soukup, T., Hazeu, G., and Jaffrain, G. (2016). *European landscape dynamics*.

[Gwenzi et al., 2016] Gwenzi, D., Lefsky, M. A., Suchdeo, V. P., and Harding, D. J. (2016). Prospects of the ICESat-2 laser altimetry mission for savanna ecosystem structural studies based on airborne simulation data. *ISPRS Journal of Photogrammetry and Remote Sensing*.

[Hazeu, 2013] Hazeu, G. (2013). *Operational land cover and land use mapping in the Netherlands*, pages 283–296. Number 18 in Remote Sensing and Digital Image Processing. Springer.

[Hazeu, 2019] Hazeu, G. (2019). Landelijk grondgebruiksbestand nederland (lgn2018). Cited 10-09-2020 from: https://www.wur.nl/upload_mm/8/b/8/c34c1088-bb2a-4f4e-81c7-a056f5e19b2b_Hazeu_LGN2018.pdf.

[Huang and Fisher, 2014] Huang, P. X. and Fisher, R. B. (2014). Individual feature selection in each one-versus-one classifier improves multi-class svm performance. *Reference Source*.

[Kropáček et al., 2014] Kropáček, J., Neckel, N., and Bauder, A. (2014). Estimation of mass balance of the grosser Aletschgletscher, swiss alps, from ICESat laser altimetry data and digital elevation models. *Remote Sensing*.

[Kurtz and Markus, 2012] Kurtz, N. T. and Markus, T. (2012). Satellite observations of Antarctic sea ice thickness and volume. *Journal of Geophysical Research: Oceans*.

[Kwok et al., 2007] Kwok, R., Cunningham, G. F., Zwally, H. J., and Yi, D. (2007). Ice, Cloud, and land Elevation Satellite (ICESat) over Arctic sea ice: Retrieval of freeboard. *Journal of Geophysical Research: Oceans*.

[Lefsky, 2010] Lefsky, M. A. (2010). A global forest canopy height map from the moderate resolution imaging spectroradiometer and the geoscience laser altimeter system. *Geophysical Research Letters*.

[Liu et al., 2020] Liu, M., Popescu, S., and Malambo, L. (2020). Feasibility of burned area mapping based on ICESAT-2 photon counting data. *Remote Sensing*.

[Markus et al., 2017] Markus, T., Neumann, T., Martino, A., Abdalati, W., Brunt, K., Csatho, B., Farrell, S., Fricker, H., Gardner, A., Harding, D., Jasinski, M., Kwok, R., Magruder, L., Lubin, D., Luthcke, S., Morison, J., Nelson, R., Neuenschwander, A., Palm, S., Popescu, S., Shum, C. K., Schutz, B. E., Smith, B., Yang, Y., and Zwally, J. (2017). The Ice, Cloud, and land Elevation Satellite-2 (ICESat-2): Science requirements, concept, and implementation. *Remote Sensing of Environment*.

[Molijn et al., 2011] Molijn, R. A., Lindenbergh, R. C., and Gunter, B. C. (2011). ICESat laser full waveform analysis for the classification of land cover types over the cryosphere. *International Journal of Remote Sensing*.

[NASA, 2018] NASA (2018). IceBridge. Cited 08-09-2020 from: https://www.nasa.gov/mission_pages/icebridge/mission/index.html.

[NASA, 2019] NASA (2019). Early Adopters. Cited 27-05-2020 from: https://icesat-2.gsfc.nasa.gov/early_adopters.

[NERC Environmental Information Data Centre, 2019] NERC Environmental Information Data Centre (2019). Land Cover Map 2015 Dataset documentation Version 1.2. Cited 07-08-2020 from: https://www.ceh.ac.uk/sites/default/files/LCM2015_Dataset_Documentation_22May2017.pdf.

[Neuenschwander and Magruder, 2019] Neuenschwander, A. L. and Magruder, L. A. (2019). Canopy and Terrain Height Retrievals with ICESat-2: A First Look. *Remote Sensing*.

[Neumann et al., 2018] Neumann, T., Brenner, A., Hancock, D., Robbins, J., Saba, J., and Harbeck, K. (2018). Algorithm Theoretical Basis Document (ATBD) for Global Geolocated Photons, ATL03. Cited 27-05-2019 from: https://icesat-2.gsfc.nasa.gov/sites/default/files/page_files/ICESat2_ATL03_ATBD_r001.pdf.

[Ni et al., 2017] Ni, H., Lin, X., and Zhang, J. (2017). Classification of ALS point cloud with improved point cloud segmentation and random forests. *Remote Sensing*.

[NSIDC, 2019] NSIDC (2019). ATLAS/ICESat-2 L3A Global Geolocated Photon Data, Version 1. Cited 05-06-2019 from: https://nsidc.org/data/ATL03/versions/1.

[NSIDC, 2020] NSIDC (2020). ICESat-2 Product Descriptions. Cited 26-03-2020 from: https://nsidc.org/data/icesat-2/products/.

[Palm et al., 2020] Palm, S., Yang, Y., and Herzfeld, U. (2020). ICESat-2 Algorithm Theoretical Basis Document (ATBD) for Global Geolocated Photons, ATL03. Cited 02-05-2020 from: https://icesat-2.gsfc.nasa.gov/sites/default/files/page_files/ICESat2_ATL04_ATL09_ATBD_r003.pdf.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Pritchard et al., 2009] Pritchard, H. D., Arthern, R. J., Vaughan, D. G., and Edwards, L. A. (2009). Extensive dynamic thinning on the margins of the Greenland and Antarctic ice sheets. *Nature*.

[scikit-learn, 2020] scikit-learn (2020). RandomForestClassifier. Cited 07-03-2020 from: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[Simard et al., 2011] Simard, M., Pinto, N., Fisher, J. B., and Baccini, A. (2011). Mapping forest canopy height globally with spaceborne lidar. *Journal of Geophysical Research: Biogeosciences*.

[Weinmann et al., 2015] Weinmann, M., Jutzi, B., Hinz, S., and Mallet, C. (2015). Semantic point cloud interpretation based on optimal neighborhoods, relevant features and efficient classifiers. *ISPRS Journal of Photogrammetry and Remote Sensing*.

[Zhang et al., 2020] Zhang, Z., Xu, N., Ma, Y., Liu, X., Zhang, W., and Li, S. (2020). Land and snow-covered area classification method based on the background noise for satellite photon-counting laser altimeters. *Optics Express*.

[Zhou et al., 2016] Zhou, Y., Qiu, F., Ni, F., Lou, Y., Zhang, C., Alfarhan, M., and Al-Dosari, A. A. (2016). Curve matching approaches to waveform classification: a case study using ICESat data. *GIScience and Remote Sensing*.

[Zwillinger and Kokoska, 1999] Zwillinger, D. and Kokoska, S. (1999). *CRC Standard Probability and Statistics Tables and Formulae*.

A. Histogram Peak Detection Algorithm

```
class Peak:
    def __init__(self, startind):
        self.born = self.left = self.right = startind
        self.died = None
    def get_persistence(self, histogram_counts):
        return float("inf") if self.died is None else histogram_counts[self.born]\
            - histogram_counts[self.died]

def histogram_peak_detection(histogram_counts):
    # The histogram counts of the histogram made from
    # the photon heights of a classification cell.
    peaks = []
    # Maps indicies to peaks
    idxtopeak = [None for s in histogram_counts]
    # Sequence indicies sorted by values
    indicies = range(len(histogram_counts))
    # This returns the ordered indicies, with the highest height first
    indicies = sorted(indicies, key = lambda i: histogram_counts[i], reverse = True)
    # Process each histogram_count in descending order
    for ind in indicies:
        lftdone = (ind > 0 and idxtopeak[ind-1] is not None)
        rgtdone = (ind < len(histogram_counts)-1 and idxtopeak[ind+1] is not None)
        il = idxtopeak[ind-1] if lftdone else None
        ir = idxtopeak[ind+1] if rgtdone else None

        # New peak (island) born
        if not lftdone and not rgtdone:
            peaks.append(Peak(ind))
            idxtopeak[ind] = len(peaks)-1
        # Directly merge to next peak/island left
        if lftdone and not rgtdone:
            peaks[il].right += 1
            idxtopeak[ind] = il
        # Directly merge to next peak/island right
        if not lftdone and rgtdone:
            peaks[ir].left -= 1
            idxtopeak[ind] = ir
        # Merge left and right peaks/islands
        if lftdone and rgtdone:
            # Left was born earlier: merge right to left
            if histogram_counts[peaks[il].born] > histogram_counts[peaks[ir].born]:
                peaks[ir].died = ind
                peaks[il].right = peaks[ir].right
                idxtopeak[peaks[il].right] = idxtopeak[ind] = il
            else:
                peaks[il].died = ind
                peaks[ir].left = peaks[il].left
                idxtopeak[peaks[ir].left] = idxtopeak[ind] = ir
    return sorted(peaks, key=lambda p: p.get_persistence(histogram_counts),\
                  reverse=True)
```

B. Feature Importance

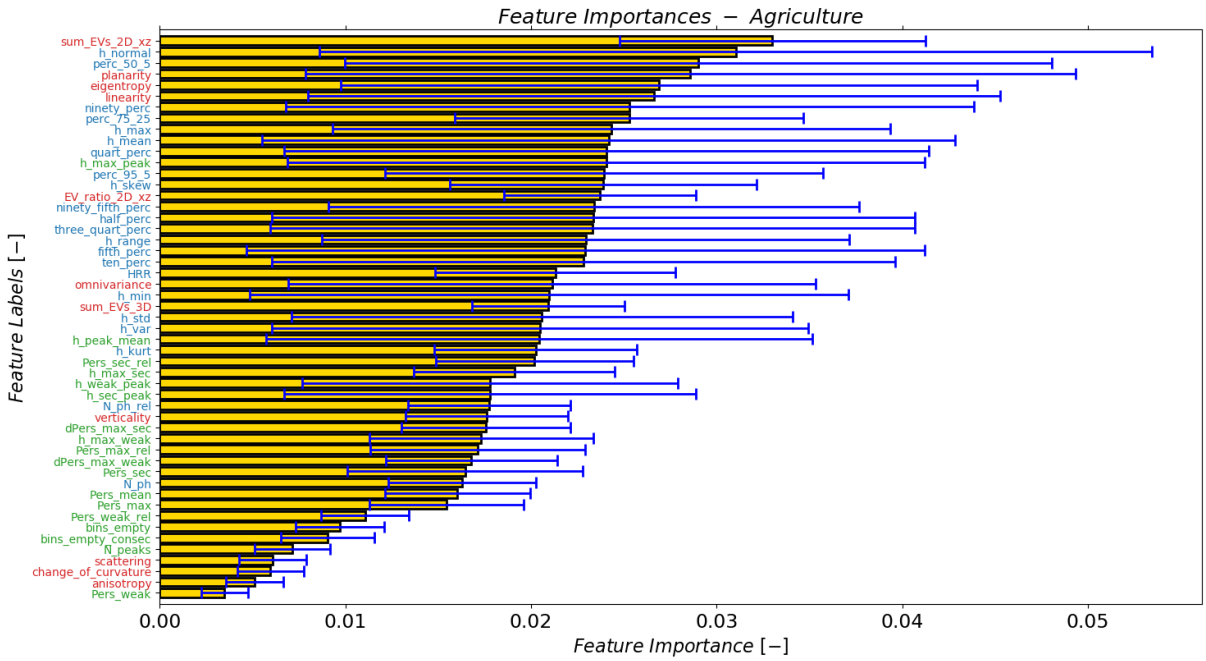


Figure B.1: Feature Importance - One vs All Classification Strategy - Agricultural Areas
51 features: 22 Altitude-derived, 18 Histogram-based, 11 Eigenvalue-based

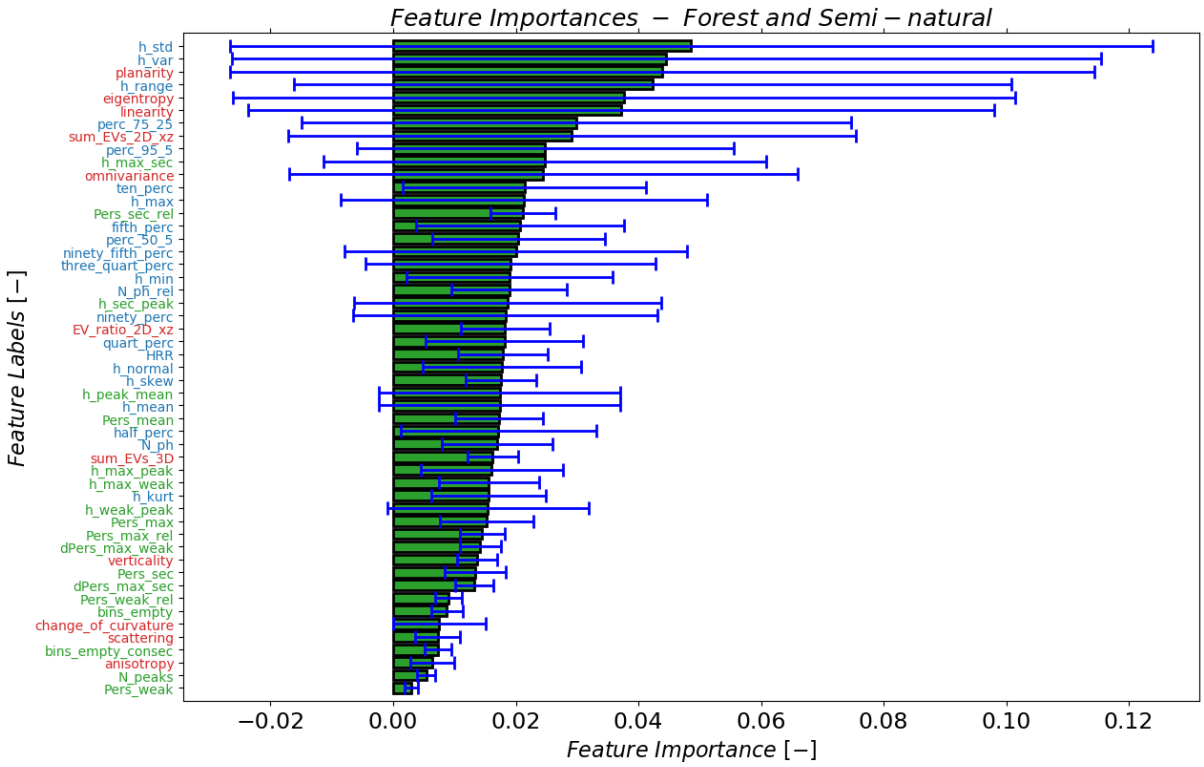


Figure B.2: Feature Importance - One vs All Classification Strategy - Forest and Semi-natural areas
51 features: 22 Altitude-derived, 18 Histogram-based, 11 Eigenvalue-based

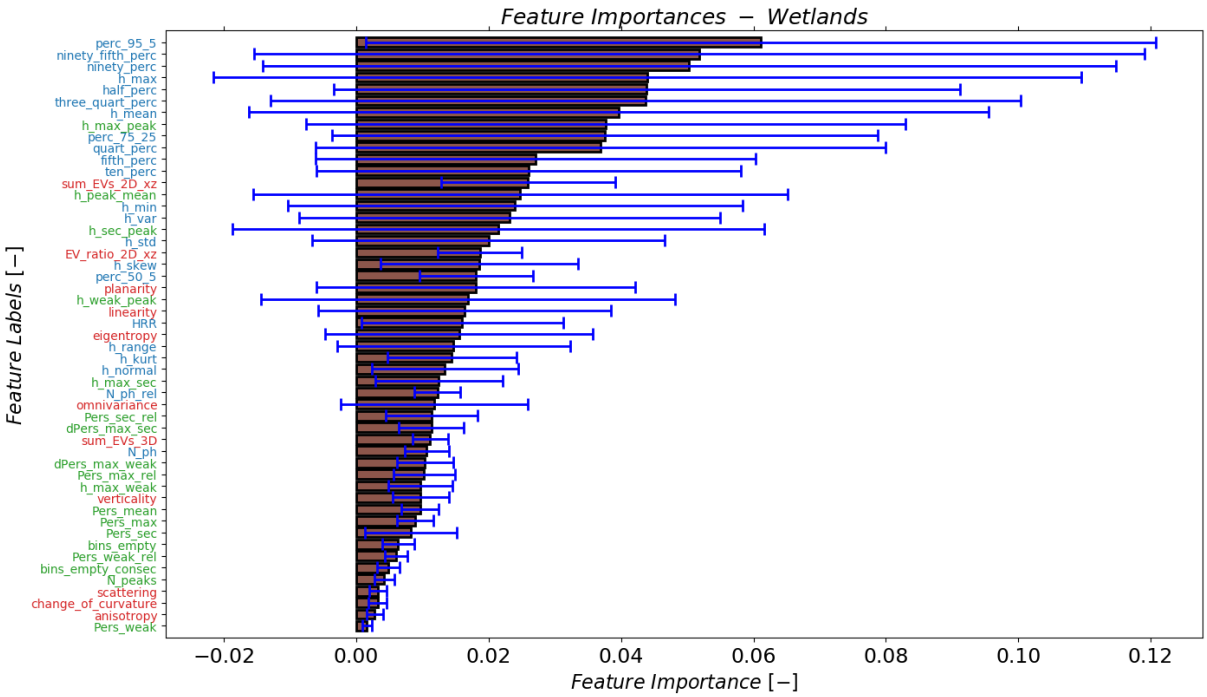


Figure B.3: Feature Importance - One vs All Classification Strategy - Wetlands
51 features: 22 Altitude-derived, 18 Histogram-based, 11 Eigenvalue-based

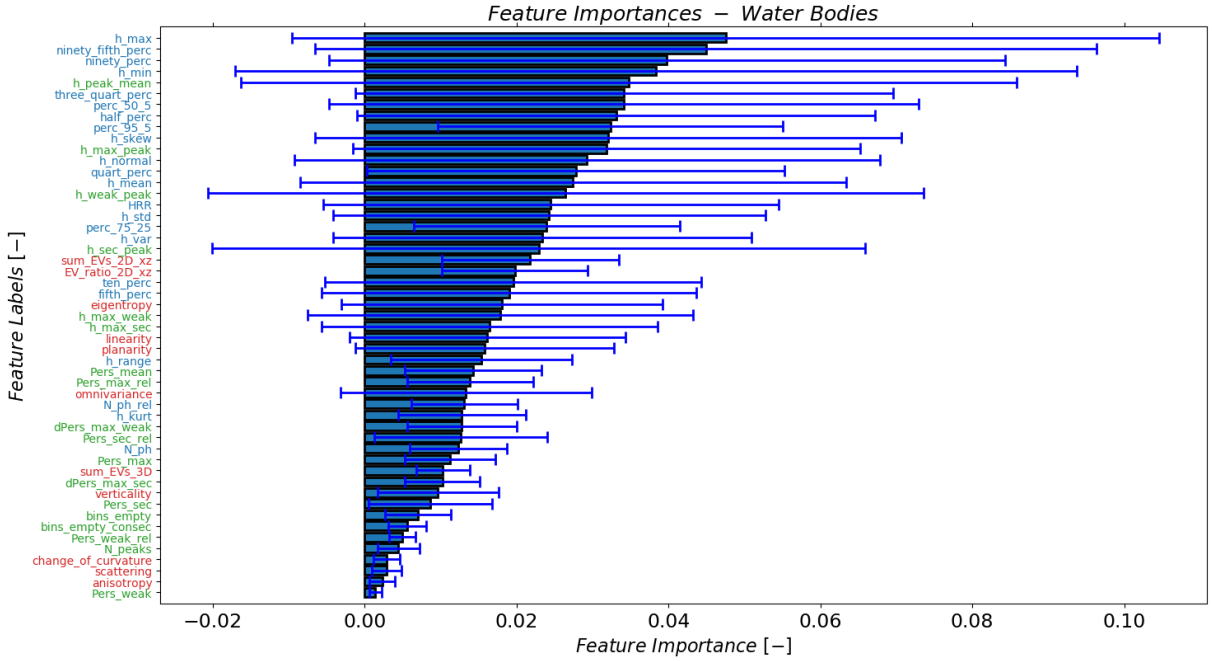


Figure B.4: Feature Importance - One vs All Classification Strategy - Water Bodies
51 features: 22 Altitude-derived, 18 Histogram-based, 11 Eigenvalue-based