Spatial Height Prediction of ICESat-2 Data using Random Forest Regression

Leo Kan
Student #5505801
1st supervisor: Hugo Ledoux
2nd supervisor: Maarten Pronk
16 June 2023

Contents

1 Introduction

In digital elevation modelling, there are many techniques for modelling terrain, such as Inverse Distance Weighted (IDW) Interpolation, kriging, and other linear spatial interpolation techniques. It is common to have a limited number of measurements and observations within the region of interest. Therefore, spatial interpolation methods are used to estimate values within the data gaps where measurements are not available. This research is going to explore how would random forest machine learning algorithm improve the existing techniques, particularly when space-borne LiDAR datasets are sparse. Some terms are used interchangeably to represent the 'bare earth' model that measures from the vertical datum. For the purpose of this research, Digital Elevation Model (DEM) will be used mainly for the bare earth model.

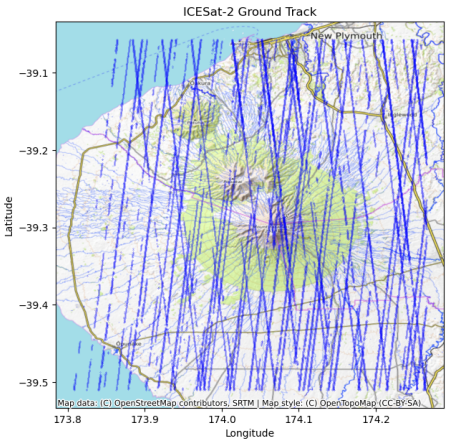


Figure 1: Cropped section of ICESat-2 Ground Track

Space-borne satellite missions like ICESat-2[1] and GEDI[2] are launched to detect changes and measure the Earth's land, ice, and vegetation surfaces with high precision. The datasets from these missions allow scientists to monitor changes in the Earth throughout its many orbits around Earth. The satellite orbits at an angle at certain intervals, and a ground track pattern is designed to cover the earth as much as possible. The ground track pattern of strong beams and weak beams of ICESat-2 is illustrated in Figure 1. More information on the ICESat-2 mission can be seen in subsection 2.1.

The altimetry data from the ICESat-2 mission will be used in this thesis to produce a digital terrain model (DTM) of several places, namely New Zealand, the Netherlands and the USA. This thesis aims to find the minimum number of features to interpolate the sparse measurements of the ICESat-2 mission using Random Forest as the main Machine Learning Algorithm, and to compare the results against the ground truth using DEM data from each of the respective mapping agencies.

Study Area and Data Sources

Three areas will be the main study area for comparison for this thesis. They are chosen to show the range of terrain features the random forest algorithm is expected to handle, namely hill, saddle, valley, ridge, and depression. The areas chosen are as follows:

- Mount Taranaki, New Zealand
- South Limburg, Netherlands
- Grand Canyon, USA

Mount Taranaki is a volcano situated in North Island, New Zealand. It demonstrates one large hill (in this case a volcano) with a height of over 2500 metres. By contrast, South Limburg, Netherlands has more terrain features with the highest peak at 300 metres. Grand Canyon, USA can show a variety of terrain features within the area.

[1] Ice, Cloud and land Elevation Satellite
[2] Global Ecosystem Dynamics Investigation

ICESat-2 data will come from the NASA Earthdata (www.earthdata.nasa.gov), and the ground truth data from the mapping agency for each country. They are Land Information New Zealand (LINZ) (www.linz.govt.nz), Het Kadaster (www.kadaster.nl, and USGS (www.usgs.gov) respectively.

2 Background and Related Work

The terrain is the surface of the Earth. This thesis is going to model an area of Earth using spatial interpolation, and the model is a representation of a 2-dimensional surface in a 3-dimensional space. Since the Earth is round, the terrain model would represent poorly on a large scale. (de Berg et al., 2008) Hence, a smaller-scale representation will be implemented. There are many spatial interpolation methods, and each has their characteristic. This section will highlight existing works on spatial interpolation and their results and accuracy metrics.

2.1 Technical Information about ICESat-2

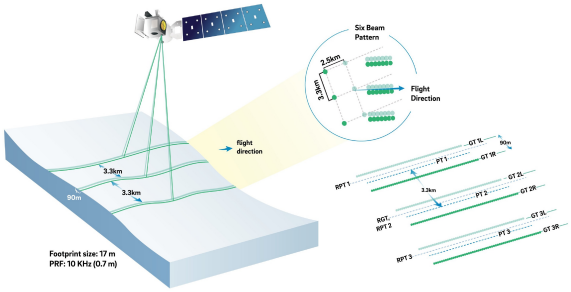


Figure 2: ICESat-2 mission beam pattern (Smith et al., 2019)

ICESat-2 is a satellite launched by NASA (2023) to monitor the land, sea, and ice elevation measurement in the Earth's polar regions. The satellite's orbit covers most of the Earth's surface due to its orbital inclination of 92 degrees, with global coverage from 88 South to 88 North latitudes. (Neumann et al., 2019) Its ground-track orbits around the polar region. The satellite uses laser beams to measure Earth's surface, and it uses three beam pairs (one weak and one string beam in each pair) that are directed to the Earth's surface to measure the elevation of the Earth's surface. This thesis will focus on data from ATL08 product [3] which has data on the along-track heights above the WGS84 ellipsoid on the ground and canopy surfaces.

The ICESat-2 operates at a pulse of 10 kHz, which means that the laser fires 10,000 times per second. This high repetition rate enables dense sampling of the Earth's surface and allows for accurate measurements of surface elevation changes. As seen from Figure 2, the gap between each track between each strong-weak beams is 3.3 kilometres and around 90 metres along each track. The ground track follows a near-polar orbit that completes the orbit around the Earth in 90 minutes. The satellite repeats itself every 91 days that covers the same ground track, and the data from repeated ground tracks enables temporal data for monitoring of changes of land, sea and ice on Earth. (Neumann et al., 2019)

Furthermore, the density of data points varies depending on the latitude of the orbit. In high-latitude polar regions, for instance, ground tracks are closely spaced, resulting in a

[3] Land and Vegetation Height Product. url: https://nsidc.org/data/atl08/versions/5

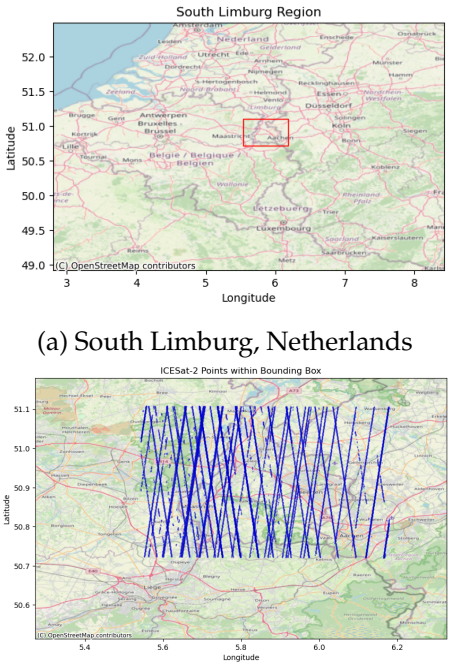
higher density of measurement points, whereas at mid-latitude and at the equator results are wider in spacing and thus lower the density of measurement points.

For the extent of this research, all points from the ATL08 data product will be used from the beginning of the satellite mission up until May 2023 within a predefined bounding box. In terms of the density of points, three areas will be chosen with similar sizes and density of points in order to find the interpolation method for such density of ICESat-2 data points.

In terms of the density and sparseness of the measurement points, this thesis aims to utilise ICESat-2 data and apply Random Forest Regression to fill in the missing areas represented as a digital elevation model raster. Specifically, this thesis will use a range of features to train the model and then evaluate its performance in predicting the missing areas, with bounding boxes having a similar density of measurements.

2.2 Existing Spatial Interpolation Methods

There are different types of interpolation methods to estimate the terrain height between the measurement points and the target points using a deterministic approach or geostatistical approach. The deterministic approach takes account of the properties of the measurement points and their neighbourhood to determine the height of the target points—these include Inverse Distance Weighting (IDW), Triangulated Irregular Network (TIN), and splines. The geostatistical approach takes account of the entire dataset to find their spatial autocorrelation—these include simple kriging and ordinary kriging.



(a) South Limburg, Netherlands

(b) ICESat-2 Points

Figure 3: Area of Interest in the Netherlands

These techniques all result in a DEM, a 2.5-dimensional representation of the terrain surface. 2.5-dimension refers to the terrain surface that only has one z value for each xy point in a 2-dimensional plain. Shapes such as overhanging structures are not permitted in 2.5D. In a DEM, surface features such as trees, buildings, and other structures are omitted, leaving the remaining surface to represent the continuous surface of the bare earth. Inverse Distance Weighting (IDW), Triangulated Irregular Network (TIN), Kriging, Splines etc. have been used extensively in previous research to produce a DEM.

Methods such as Inverse Distance Weighting (IDW), Triangulated Irregular Network (TIN), Natural Neighbour Interpolation (NNI), and Lapalce Interpolation are deterministic interpolation methods that are able to calculate points that are in between measurement points. These methods are widely in use because they are relatively simple to implement, and can result in a relatively accurate representation of the ground truth data, provided that the data points are equally distributed with sufficient density to recognise terrain features in the sampled data.

Existing deterministic interpolation techniques may not function as expected on sparse

datasets. In the case of ICESat-2 data, the region of South Limburg, Netherlands is used for the region of interest to illustrate the interpolation. The bounding box (seen in Figure 3a and Figure 3b) covers most of the south part of Limburg, Netherlands, and the border regions of Germany and Belgium.

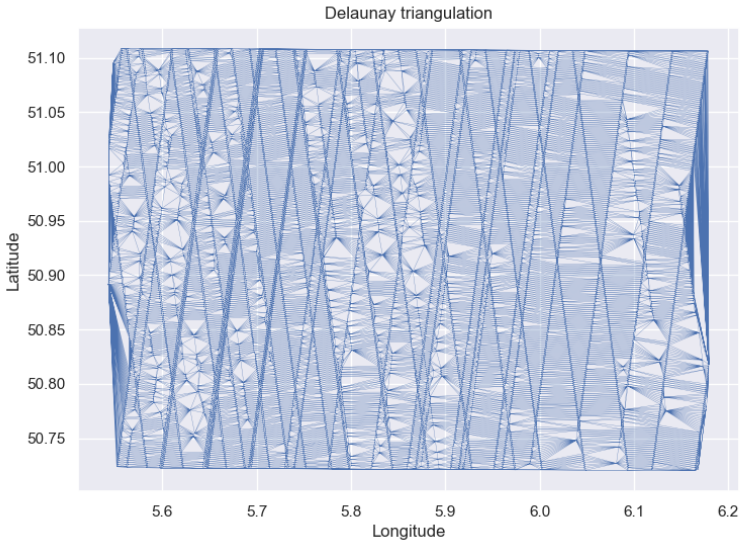


Figure 4: Delaunay triangulation

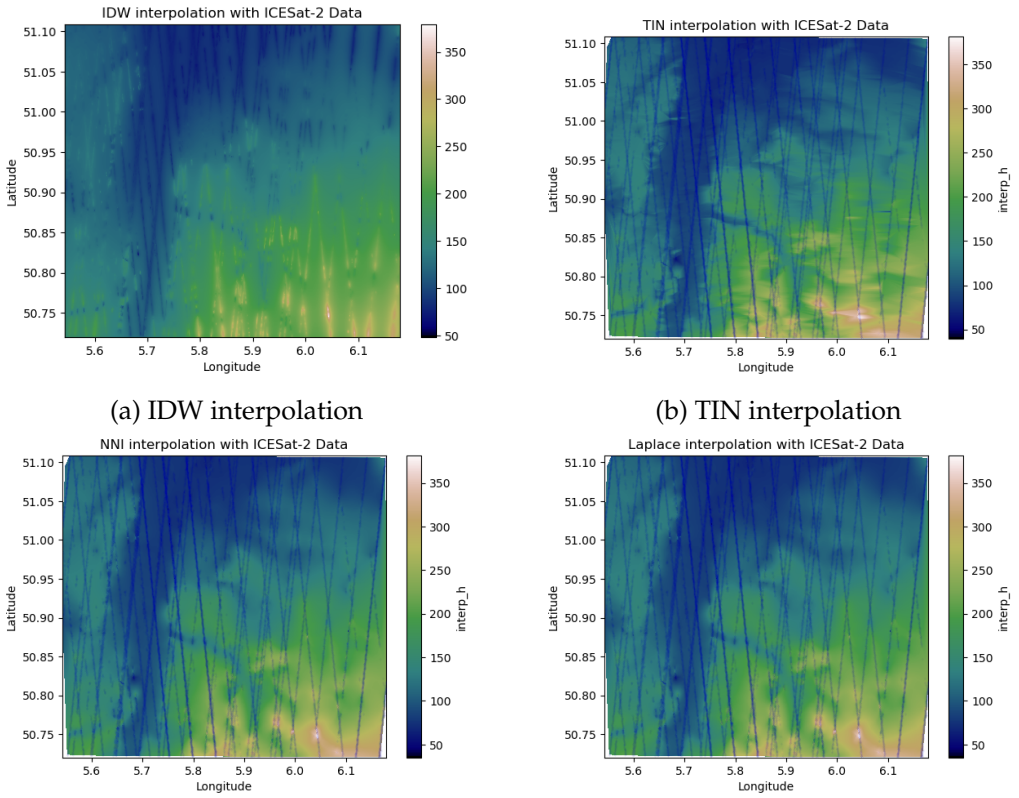
Since TIN, NNI and Laplace interpolation are built upon on Delaunay triangulation. The representation of triangles in Figure 4 shows that the triangles are constructed following the points of the ICESat-2 points with long thin triangles along the path of the satellite ground track. The shape of the ground track can also be seen clearly in the Delaunay triangulation within the convex hull as illustrated in Figure 4.

Traditional interpolation methods have their drawbacks too. Figure 5 shows different interpolation methods and their results show that some details of terrain features are smoothed and using a 30m resolution cannot be able to interpolate a good terrain feature since the space between the satellite ground track is too wide to be able to determine smaller details in between. There are techniques such as random forest algorithms to use machines to learn auxiliary datasets as features such that terrain features can also be recovered, which is discussed in subsection 2.4

Inverse Distance Weighting (IDW)

IDW puts weight to measurements that are closer to the nearby measurement points. This method is easy to implement since all data points apply a power parameter, thus the area closer to the data point is inversely proportional to the power of its distance. Closer points have a greater weighting than areas further from the data point. IDW can, however, produce sharp peaks depending on the power parameter, due to the decaying factor that influences the area surrounding the data point.

Despite its shortcomings, IDW is a popular method for interpolation in Geographic Information Systems (GIS) software packages. A comparative study between different interpolation methods, Arun (2013) observed that IDW is a good interpolation method for morphologically smooth areas as the study has kept the study area of 4 km^2.



(a) IDW interpolation

(b) TIN interpolation

(c) NNI Interpolation

(d) Laplace Interpolation

Figure 5: Traditional Interpolation Techniques

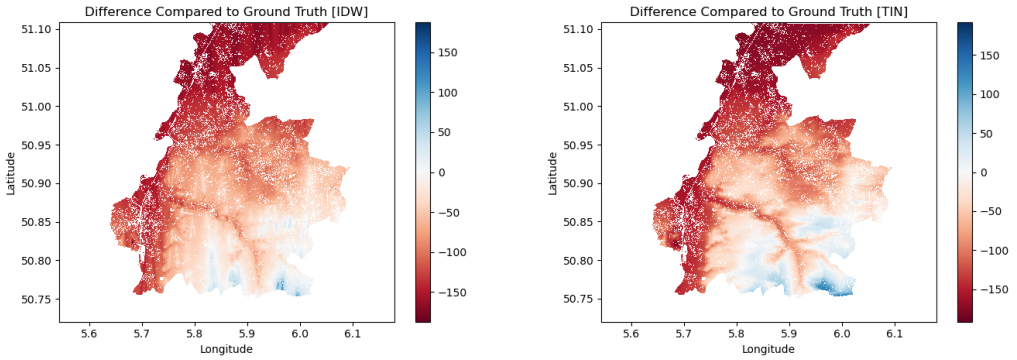
Triangulated Irregular Network (TIN)

Another popular interpolation method, TIN interpolation follows the shape of the terrain based on the data points using Delaunay Triangulation. The values are based on the network of the triangles, each vertex connecting to the data points with known heights. Each triangle represents a small portion of the terrain and the elevation values are determined within the triangles at any point within itself. This is normally represented by terrain with triangles covered throughout the terrain. TIN and Delaunay Triangulation is then also used in other interpolations such as Laplace and NNI.

Kriging

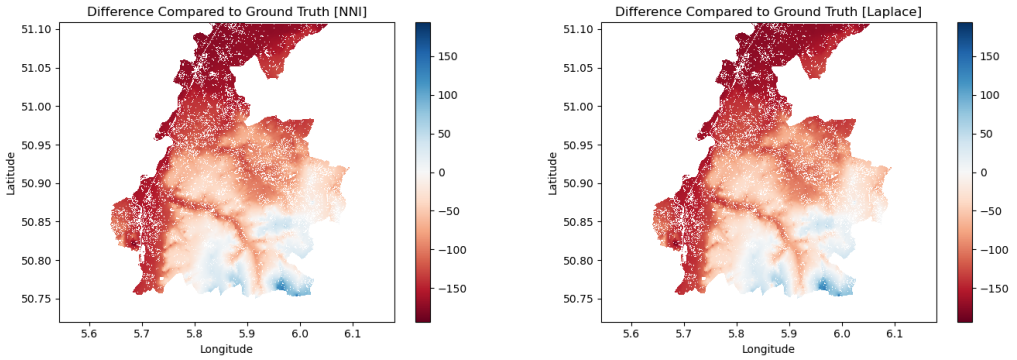
Kriging is a geostatistical method used in spatial data analysis. It predicts unknown values for a target variable at unsampled locations based on observed values at nearby sampled locations. The technique is based on the assumption that the spatial correlation between the target variable at different locations can be described by statistics. The variogram shows the spatial relationship between the point pairs–the pair of points that are closer to each other are more related. The drawback of kriging is that

For large data gaps or holes, Luedeling et al. (2007) used other external datasets to fill the void in between the SRTM (https://gedi.umd.edu/) data. Before 2010, The second version of SRTM DEM data contained voids in the mountainous terrains. The author extracted the voids into polygon and used data from Russian topographic survey maps as proxy data and filled the data gaps. Both maps were then converted to a refined TIN and voids were filled such that it has a continuous surface. The characteristic of this approach



(a) IDW interpolation

(b) TIN interpolation



(c) NNI Interpolation

(d) Laplace Interpolation

Figure 6: Traditional Interpolation against Ground Truth

however would lead to inconsistencies and borders in the gap-filled areas.

2.3 Forest And Buildings removed Copernicus DEM (FABDEM)

FABDEM is a DEM that is derived based on the Copernicus GLO-30 DEM distributed by the European Space Agency. Fundamentally, Copernicus DEM falls towards DSM, which is a terrain model that includes both natural and man-made features—meaning building and vegetation height existing within the raster dataset. FABDEM, hence, MLA aims to create a global DEM that represents the 'bare earth' version of Copernicus DEM.

The creation of FABDEM, according to Hawker et al. (2022), uses random forest regression to remove buildings and vegetation from Copernicus DEM. Key datasets such as forest cover, building footprint, and forest height, are used to remove building and vegetation separately. The creation of FABDEM, is compared against other globalDEM to maintain good accuracy.

2.4 Interpolation Methods using Machine Learning

Machine learning algorithms (MLA) in the past decade has emerged as an alternative to deterministic and geostatistic approaches to perform spatial interpolation. This is owing to the advancement in virtual machines and servers, so complex operations can be performed. Easier access to these computers also signifies that computationally expensive operations can be outsourced to a high-performance computer. Delft High Performance Computing Centre (DHPC) (2022), for instance, provides access to hardware capabilities to perform machine learning and deep learning algorithms. Recent research, to a larger

extent, now focuses on using machine learning to predict the height of the target points in DEM, as well as other fields such as soil science and geology.

The fact that ICESat-2 data is sparser at locations close to the equator means deterministic methods of interpolation are impossible to apply with the kilometre gaps. Similarly, the geostatistical approach can also be very unreliable and computationally expensive. Machine learning, therefore, is a feasible approach to develop a DEM with a selection of auxiliary data as features.

3 Research Question

This project will primarily use Machine Learning algorithm to predict the height of each pixel using auxiliary data from various sources. The raw data will be sourced from NASA EarthData, auxiliary data for machine learning training will be sourced from OpenTopography (2023) and the ground truth data sourced from New Zealand's mapping agency, LINZ. The main research question of this thesis reads:

- Is it possible that using random forest machine learning algorithm, is able to construct DEM with sparse isotropic ICSat-2 satellite data?
- What is the minimum number of features that is able to reconstruct a DEM with good accuracy?
- With 3 areas of interest of similar density of ICESat-2 data points, can the algorithm reconstruct a DEM with good accuracy?

3.1 Datasets and tools used

ICESat-2 dataset

The dataset for ICESat-2 is available from NASA Earthdata, and the ATL08 along-track data product is the main data source for the space-borne LiDAR. Neuenschwander et al. (2021) provides a quick look at the ATL08 product, and the `icepyx` python library, made available by the National Snow & Ice Data Center (NSIDC) provides the main access portal for ICESat-2 data. (Scheick, 2019)

Features Dataset

The auxiliary data will be sourced from Copernicus DEM. This dataset, however, is a DSM that is freely available from OpenTopography (2023). This will be used as auxiliary data for machine learning training. Derived datasets from Copernicus DEM such as Aspect and Slope will be used as features in the machine learning algorithm in this project. In addition, land use and land cover data will be accessed from Copernicus Land Monitoring Service for the land use information in the set of auxiliary data. The full list of features dataset can be found in Table 1

Dataset for Assessment

Since there are three regions of interest in this project, the datasets for accuracy assessment will be DEM from LINZ (New Zealand), Het Kadaster (Netherlands), and USGS (United States) respectively. These mapping agencies from their respective countries

will provide the dataset for ground truth assessment for verification of the data results.

4 Methodology

The objective for this thesis project is to model a DEM from ICESat-2 data points obtained from the ICESat-2 mission from NSIDC. To obtain the desired interpolation, random forest regression is used to model the data points. Random forest regression is the process of predicting continuous numerical values using multiple decision trees. Their outputs are then averaged in order to predict the height of the spaces between the data points. A random forest provides improved accuracy and robustness compared to individual decision trees, and this thesis will use this to model a DEM. The diagram of the workflow is illustrated in Figure 7.

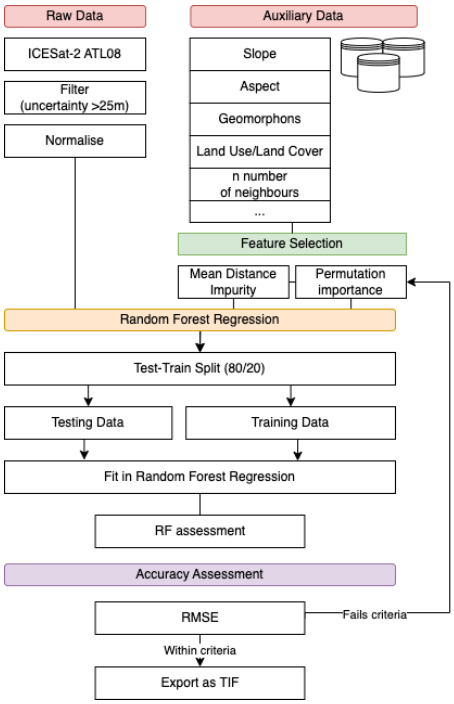
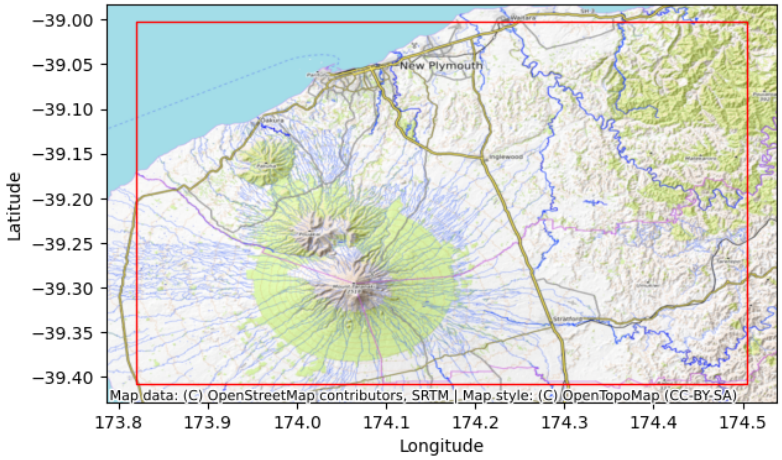


Figure 7: Workflow for this project

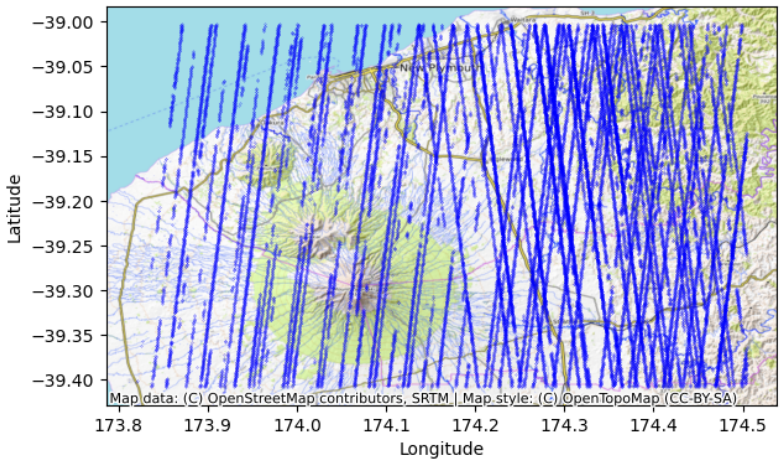
For the random forest regression model of the DEM, various auxiliary data are needed as inputs into the regression model. The dataset from Copernicus DEM, land use and land cover, slope, and aspect will be used as auxiliary data for the regression. These auxiliary data are the inputs to the random forest regression. Using the auxiliary data, the random forest will generate decision trees and produce a regression analysis, resulting in a prediction of the height for each pixel in the DEM raster.

4.1 The Region of Interest

The region of interest shown in a red box in Figure 8a is located on the North Island of New Zealand. The volcano, Mount Taranaki, is situated in the Taranaki Peninsula on the west coast of New Zealand's North Island, and the peninsula extends out into the Tasman Sea. The region of Egmont National Park provides a good range of terrain features. The lower slopes of the volcano are covered in dense forest, while the higher elevations are



(a) Region of Interest



(b) ICESat-2 transects over the ROI

Figure 8: Region of Interest

more rocky and barren. Since New Zealand is also one of the geologically active regions in the world, Mount Taranaki is chosen to be the region of interest in this research.

At the end of 2022, LINZ has uploaded a LiDAR dataset that covers this region—including DEM and DSM at 1-meter resolution and an aerial LiDAR dataset provided by LINZ that can be accessed through OpenTopography (www.opentopography.org). This 1-metre resolution DEM will be used for accuracy assessment after producing the random forest regression DEM model.

4.2 Obtaining ICESat-2 ATL08 Data

Throughout this project, Python will be the programming language along with *numpy*, *pandas*, *geopandas* etc. libraries for additional functions. The *icepyx* python library is the data portal provided from Scheick (2019) so that ICESat-2 data can be downloaded. With the *icepyx*, only HDF5 data can be downloaded with the download methods with the bounding box as input.

The downloaded HDF5 data contains a large amount of data. Each of the data points obtained from ICESat-2 mission is embedded with data including the name of the granule, each of the beams, longitude, latitude, photon rate etc. As the scope of this project fo-

cuses on terrain information, ground information is used primarily. The ALT08 land and vegetation product contains the dataset that is needed for this project. Within the ATL08 product, the estimated best fit, **'h_te_best_fit'**, which is the estimated photon ground points from the centre of each 100m step Neuenschwander and Pitts (2019), the *'latitude'* and *'longitude'* are used as the essential dataset.

4.3 Obtaining auxiliary data as feature dataset

Table 1: Auxiliary Data used in this thesis

Auxiliary Data	**Data source**
Land Based Dataset	
World Settlement Footprint	Deutsches Zentrum für Luft- und Raumfahrt (DLR)
Land Water Mask	US Geological Survey
Global Land Cover	Copernicus Global Land Service
Global High-Resolution Geomorphometric Layers	OpenTopography
Geometry Based Dataset	
100 closest ICESat-2 data to raster grid	Using KD Tree on ICESat-2 data
Inverse Distance Weighting (IDW)	Using ICESat-2 interpolation from Figure 5
Triangulated Irregular Network (TIN)	Using ICESat-2 interpolation from Figure 5
Natural Neighbour Interpolation (NNI)	Using ICESat-2 interpolation from Figure 5
Laplace Interpolation	Using ICESat-2 interpolation from Figure 5

Random forest regression requires auxiliary data for its prediction. The main data source will come from Copernicus DEM, as shown in Table 1. Copernicus DEM is a high resolution map of the Earth's surface elevation–also called Digital Surface Model (DSM)–that is an edited version of the WorldDEM product. The DSM is generated using data from the TanDEM-X satellite mission in partnership with German State and Airbus. Some editing has been done to coastlines, shorelines, and irregular terrain structures to enhance the accuracy and reliability of the model.

For the purpose of this research, The 'GLO-30' series will be used to cover the region of interest. 'GLO-30' is a raster that represents 30m resolution of the DSM product that is near global coverage. This raster data will be the base computation of other auxiliary data such as slope and aspect. The benefit of using this raster is that it is freely available to download through the Copernicus Open Access Hub (European Space Agency, 2022). The auxiliary data will also be the 100 closest ICESat-2 data point, by distance, between each of the ICESat-2 data; and the 100 closest datasets between the raster grid mid-points and the ICESat-2 data point.

4.4 Filtering Data

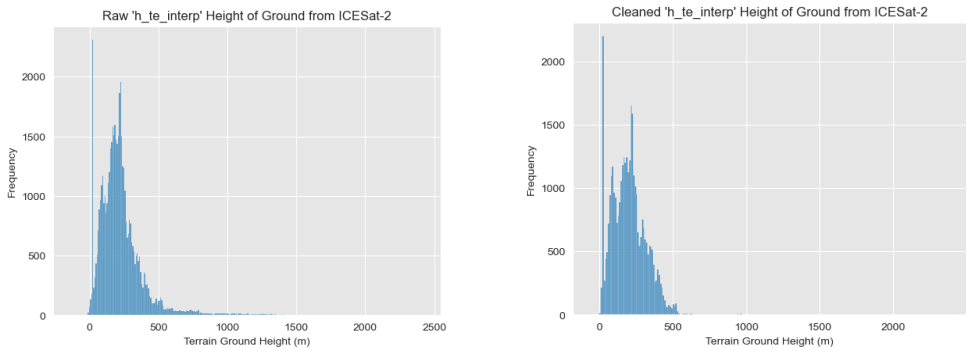
	h_te_interp	**h_te_uncertainty**
count	49796	49796
mean	233.108	7.452e+37
std	194.729	1.407e+38
min	-134.360	0.013
25%	129.822	1.649
50%	201.958	5.162
75%	278.744	39.867
max	2421.418	3.403e+38

(a) Raw ICESat-2 Data

	h_te_interp	**h_te_uncertainty**
count	36019	36019
mean	205.772	4.677
std	136.335	5.119
min	-55.160	0.013
25%	111.958	1.192
50%	195.907	2.677
75%	270.026	6.274
max	2374.881	24.991

(b) Cleaned ICESat-2 Data

Table 2: Statistics of ICESat-2 Data



(a) Raw Data of **h_te_interp** (b) Cleaned Data of **h_te_interp**

Figure 9: Histogram of ICESat-2 Data

With such a large amount of data embedded in HDF5 format, there are a sizeable amount of data that needed filtering down. In terms of the data, that are data that are considered essential to this research, namely, the latitude and longitude and the best-fit height based on the WGS84 ellipsoid. Additionally, the uncertainty of each ICESat-2 data point is also needed to assess the quality of the data. Table 2a provided some basic statistics of the raw data captured by the satellite and Table 2b shows the cleaned data.

As part of data cleaning, an assessment of data quality is required. This step involves filtering out data points that are considered outliers. For height data, Table 2a shows the height (**h_te_interp**) minimum at -134.36 metres and maximum at 2,421.42 metres. The raw data with a negative height of 100 metres is considered unreasonable, especially when the standard deviation of uncertain at such an unrealistically large magnitude. Cleaning of the data, therefore, is a necessary step to ensure the data input into the random forest has a good range.

With reference to the Algorithm Theoretical Basis Document (ATBD) of the ATL08 product, the **h_te_uncertainty** parameter is the total uncertainly of ground height estimates that includes "uncertainties such as geo-location, pointing angle, timing, radial orbit error" Neuenschwander et al. (2021). In the case of the ICESat-2 raw data for the region of interest, **h_te_uncertainty** of more than 25 metres will be deleted, and outliers of more than 3 standard deviations will also be deleted from the raw dataset.

4.5 Data normalisation

$$h_{normalised} = \frac{h - \min(h)}{\max(h) - \min(h)}$$

Data normalisation is to constrain the elevation values of the terrain to a consistent scale or range. Since three areas of interest have different ranges in height, normalisation of height data to the ICESat-2 dataset will be performed such that the scale of the **h_te_interp** data is within the range where the minimum value will be 0 and maximum value will be 1.

4.6 Feature Selection and Ranking

The goal of feature selection is to identify the most informative and influential features that contribute the most to the predictive performance of a model or the interpretability of the data. The use of feature selection and ranking is to determine which of the features in

the machine learning algorithm produces the most important and relevant dataset. Out of the auxiliary datasets in **??**, using feature selection and ranking can help determine which of the feature datasets have a greater impact on the model's performance.

4.7 Random Forest Regression

The main toolkit used for random forest regression in this research will be to use the Python library *scikit-learn*. *Scikit-learn* is a machine learning library for Python that supports both supervised and unsupervised machine learning. Using this library, random forest regression of the terrain data can be placed in the random forest algorithm to predict the height value as a DEM. Using random forest regression as a spatial predictions framework (RFsp) was based on the work by Hengl et al. (2018). Random forest data will take into account the spatial correlation of the data points and the raster points of the resulting DEM. Sekulić et al. (2020) added that the RFsp framework can work spatial data and its correlation. These auxiliary data can improve the prediction and the predicted result that is similar to kriging.

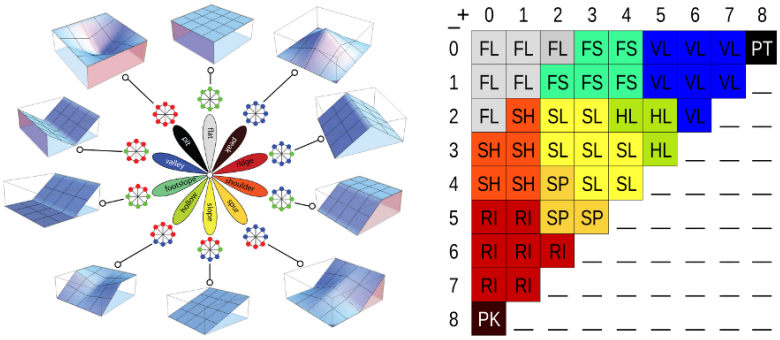


Figure 10: Geomorphons (Jasiewicz and Stepinski, 2013)

The auxiliary data used in this thesis will also incorporate various sources like. Geomorphons are classifications of landforms in the raster data (Jasiewicz and Stepinski, 2013). Each of the terrain features is classified based on the neighbouring cells. The classification, illustrated in Figure 10, is divided into 10 landforms representing flat, peak, ridge, shoulder, spur, slope, hollow, footslope, valley and pit. The resulting raster file will then be used as the auxiliary for the random forest algorithm.

4.8 Random Forest Algorithm

Random forest will use a lot of decision trees to decide the height of the terrain using aggregation and bagging. The majority voting from the multiple trees from the result of aggregation will determine the height of the spaces between the measured points.

The incorporation of auxiliary data as features during the training data means that the parameters of each decision tree will be based on the feature data to decide whether the decision goes left or right of the decision tree. Due to the data process, the resulting data should produce a more accurate, closer result of the ground truth terrain.

In terms of the machine learning algorithms, there will be constrained to the Scikit Learn library since it is supported by Python and also its ease of implementation of different machine learning algorithms.

4.9 Accuracy Assessment of Random Forest Regression

The random forest regression is comprised of the input of ICESat-2 data and an array representing the raster. In order to ascertain the accuracy of random forest regression, a test train split is done to ensure the training data and testing data perform within the constraints of the machine learning algorithm. Mean Square Error and R^2 are common assessment criteria to perform accuracy assessment, and they are within the Scikit-learn library. It is easily accessible within the library itself.

4.10 Accuracy Assessment of Resulting DEM Raster

In order to compare the results against the ground truth, there are assessment methods that are used in this project. Firstly, due to the sparseness of the ICESat-2 dataset, the resolution of the ground truth raster file will match the resolution of the raster file to around 500m resolution. The comparison assessments that will be used are Median Absolute Error (MAE) and Root Mean Square Error (RMSE). Both MAE and RMSE calculate the magnitude of the differences between the ground truth and resulting DEM.

Mean Absolute Error (MAE)

$$\text{MAE} = \frac{\sum_{i=1}^{n} |h_i - \hat{h}_i|}{n}$$

MAE is a measure of the average absolute difference between the height values of the ground truth compared to the predicted height. This measurement is less sensitive to outliers and is always a non-negative number. The lower value for MAE indicates higher accuracy with the best being 0.

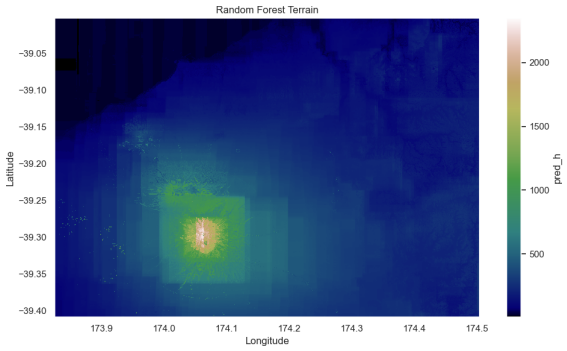
Root Mean Square Error

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (h_i - \hat{h}_i)^2}{n}}$$

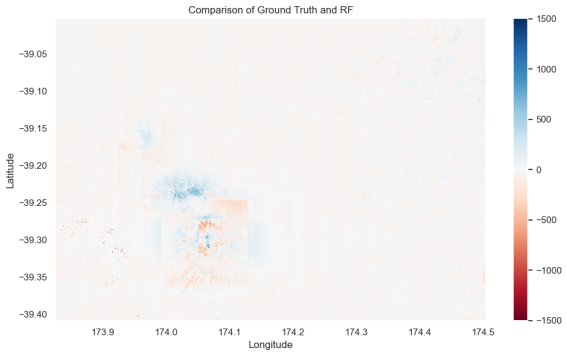
RMSE is a measure of the squared differences between the height of the resulting DEM and the ground truth. It is used widely in terrain analysis. The lower value for MAE indicates indicate higher accuracy and a better correlation between the resulting DEM and the ground truth.

5 Preliminary Results

Preliminary results in the random forest use Mount Taranaki as the region of interest, and the model predicted has some differences between it and the ground truth. It shows some straight lines that might be due to the decision trees making their decisions. The range of deviation is quite large with the highest deviation from the ground truth between -866m to 1003m.



(a) DEM generated from Random Forest



(b) Comparison to Grounth Truth

Figure 11: Preliminary Results of Random Forest Regression

The features used in the random forest regression are slope, aspect, roughness, and land cover. The resulting model has some large deviations, but the peak of the volcano can still be observed and illustrated in Figure 11 and the 3D visualised image in Figure 12

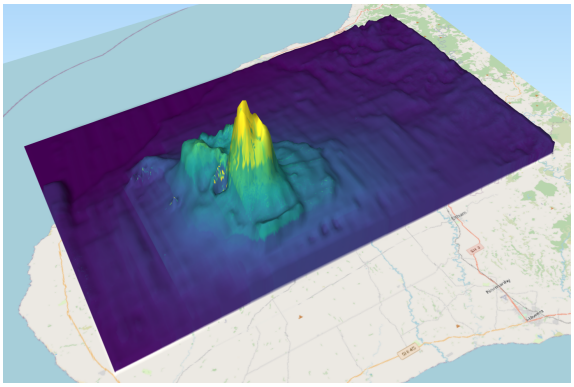


Figure 12: Random Forest Regression DEM

6 Time Planning

The Gantt chart below highlights the task that will be performed throughout the thesis writing period.

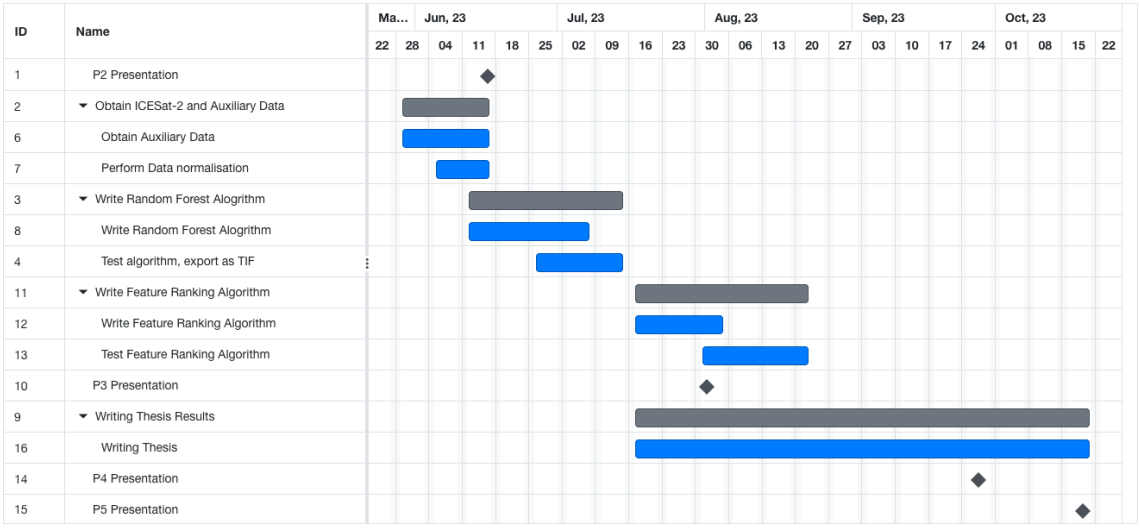


Figure 13: Grantt Chart

7 Bibliography

P. V. Arun. A comparative analysis of different DEM interpolation methods. *The Egyptian Journal of Remote Sensing and Space Science*, 16(2):133–139, Dec. 2013. ISSN 1110-9823. doi: 10.1016/j.ejrs.2013.09.001. URL https://www.sciencedirect.com/science/article/pii/S1110982313000276.

M. de Berg, O. Cheong, M. Van Kreveld, and M. Overmars. Delaunay Triangulations: Height Interpolation. *Computational Geometry*, pages 191–218, 2008. doi: 10.1007/978-3-540-77974-2_9. URL http://link.springer.com/10.1007/978-3-540-77974-2_9.

Delft High Performance Computing Centre (DHPC). DelftBlue Supercomputer (Phase 1), 2022. URL https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase1.

European Space Agency. Copernicus DEM, 2022. URL https://spacedata.copernicus.eu/collections/copernicus-digital-elevation-model. Institution: European Space Agency.

L. Hawker, P. Uhe, L. Paulo, J. Sosa, J. Savage, C. Sampson, and J. Neal. A 30 m global map of elevation with forests and buildings removed. *Environmental Research Letters*, 17(2):024016, Feb. 2022. ISSN 1748-9326. doi: 10.1088/1748-9326/ac4d4f. URL https://dx.doi.org/10.1088/1748-9326/ac4d4f. Publisher: IOP Publishing.

T. Hengl, M. Nussbaum, M. N. Wright, G. B. Heuvelink, and B. Gräler. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, 6:e5518, Aug. 2018. ISSN 2167-8359. doi: 10.7717/peerj.5518. URL https://peerj.com/articles/5518.

J. Jasiewicz and T. F. Stepinski. Geomorphons — a pattern recognition approach to classification and mapping of landforms. *Geomorphology*, 182:147–156, Jan. 2013. ISSN 0169555X. doi: 10.1016/j.geomorph.2012.11.005. URL https://linkinghub.elsevier.com/retrieve/pii/S0169555X12005028.

E. Luedeling, S. Siebert, and A. Buerkert. Filling the voids in the SRTM elevation model — A TIN-based delta surface approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 62(4):283–294, Sept. 2007. ISSN 09242716. doi: 10.1016/j.isprsjprs.2007.05.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0924271607000433.

NASA. ICESat-2, Jan. 2023. URL https://icesat-2.gsfc.nasa.gov/.

A. Neuenschwander and K. Pitts. The ATL08 land and vegetation product for the ICESat-2 Mission. *Remote Sensing of Environment*, 221:247–259, Feb. 2019. ISSN 0034-4257. doi: 10.1016/j.rse.2018.11.005. URL https://www.sciencedirect.com/science/article/pii/S0034425718305066.

A. L. Neuenschwander, K. L. Pitts, B. P. Jelley, J. Robbins, B. Klotz, S. C. Popescu, R. F. Nelson, D. Harding, D. Pederson, and R. Sheridan. ATLAS/ICESat-2 L3A Land and Vegetation Height, version 5, 2021. URL http://nsidc.org/data/atl08/versions/5.

T. A. Neumann, A. J. Martino, T. Markus, S. Bae, M. R. Bock, A. C. Brenner, K. M. Brunt, J. Cavanaugh, S. T. Fernandes, D. W. Hancock, K. Harbeck, J. Lee, N. T. Kurtz, P. J. Luers, S. B. Luthcke, L. Magruder, T. A. Pennington, L. Ramos-Izquierdo, T. Rebold, J. Skoog, and T. C. Thomas. The Ice, Cloud, and Land Elevation Satellite – 2 Mission: A Global Geolocated Photon Product Derived From the Advanced Topographic Laser Altimeter System. *Remote sensing of environment*, 233:10.1016/j.rse.2019.111325, Nov.

2019. ISSN 0034-4257. doi: 10.1016/j.rse.2019.111325. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6839705/.

OpenTopography. OpenTopography, Jan. 2023. URL https://opentopography.org/.

J. Scheick. icepyx: Python tools for obtaining and working with ICESat-2 data, 2019. URL https://github.com/icesat2py/icepyx.

A. Sekulić, M. Kilibarda, G. B. M. Heuvelink, M. Nikolić, and B. Bajat. Random Forest Spatial Interpolation. *Remote Sensing*, 12(10):1687, Jan. 2020. ISSN 2072-4292. doi: 10.3390/rs12101687. URL https://www.mdpi.com/2072-4292/12/10/1687. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

B. Smith, H. A. Fricker, N. Holschuh, A. S. Gardner, S. Adusumilli, K. M. Brunt, B. Csatho, K. Harbeck, A. Huth, T. Neumann, J. Nilsson, and M. R. Siegfried. Land ice height-retrieval algorithm for NASA's ICESat-2 photon-counting laser altimeter. *Remote Sensing of Environment*, 233:111352, Nov. 2019. ISSN 0034-4257. doi: 10.1016/j.rse.2019.111352. URL https://www.sciencedirect.com/science/article/pii/S0034425719303712.