Estimating building height from ICESat-2 data: the case of the Netherlands

Ziyan Wu
student #5360684

1st supervisor: Hugo Ledoux
2nd supervisor: Maarten Pronk
Co-reader: Dr. Azarakhsh Rafiee
June 29, 2022

Space-borne LiDAR -- ICESat2

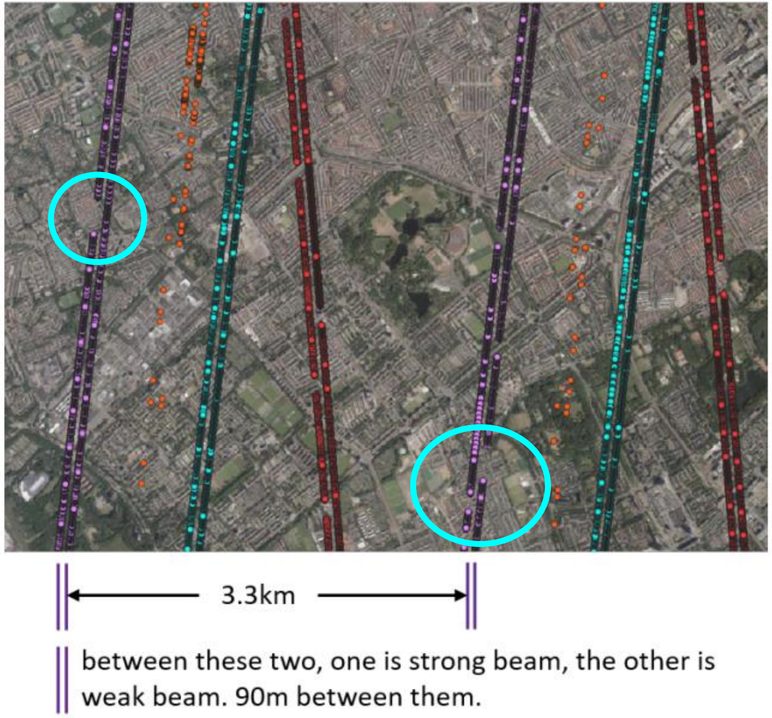
The Ice, Cloud, and land Elevation Satellite-2 (**ICESat-2**) was launched in 2018, employing photon-counting LiDAR to collect Earth's surface elevation data globally. Coverage up to 88°N–88°S latitude.

The **beam pairs** are separated by 3.3 km in the cross-track direction and the distance between the **strong and weak beam** in a pair is 90 m.

Sparsity

- Distance between beams.
- Different beams have various distribution pattern. Orange one is sparse than others.
- The photon distribution on the same beam is not evenly. Some parts are missing (blue circle in purple beam).

16 beams (8 pairs of strong and weak beams)



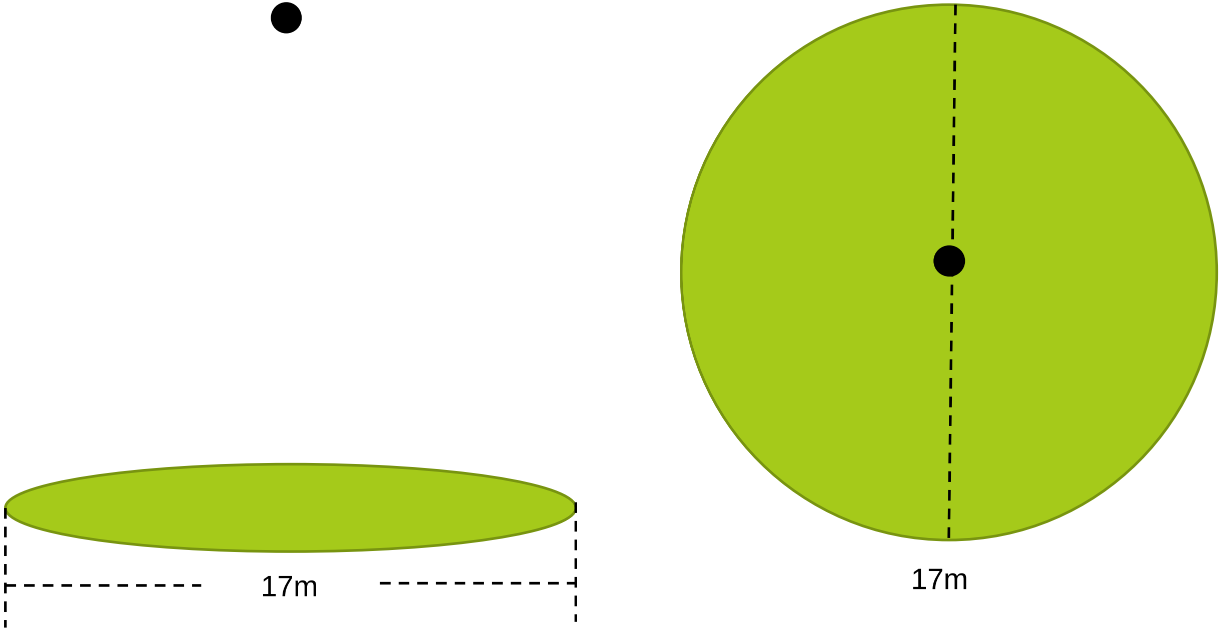
Space-borne LiDAR -- ICESat2

Footprint:

Each photon point from ICESat-2 has a footprint of approximately 17m in diameter.

And as time grows and energy decrease, this value could increase to about 20m in three years[31].

In theory, the elevation obtained by ICESat-2 point could be any objects inside this diameter.



TU Delft

[31] Neuenschwander, A. and Pitts, K. (2019). The ATL08 land and vegetation product for the ICESat-2 Mission. Remote Sensing of Environment, 221:247–259.

Research overview

Aim:

1) Get an acceptable prediction model (ML) that can be used to estimate the height of all buildings in the Netherlands

	Features				
samples	Building height	Construction year	…	…	…
s1					
…	…	…	…	…	…

Training dataset (some buildings data in NL)

train

Prediction model

Algorithms and statistical models are used to enable computer to find patterns between building height (y) and other features (x) in large amounts of data.

Then use the model can predict or describe new data.

2) Calculated building height from ICESat-2 data, explore can ICESat-2 data be used in building height retrieval area.

Data:

ICESat-2

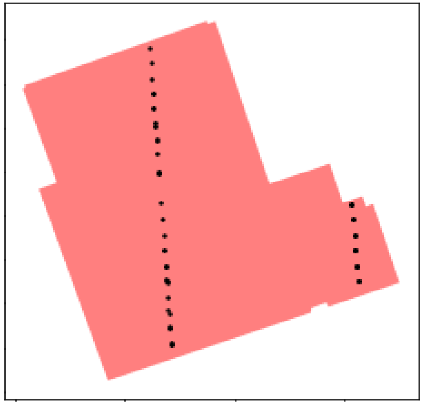
- icesat-2 point data includes ground / roof elevation
- building height can be calculated

3D BAG

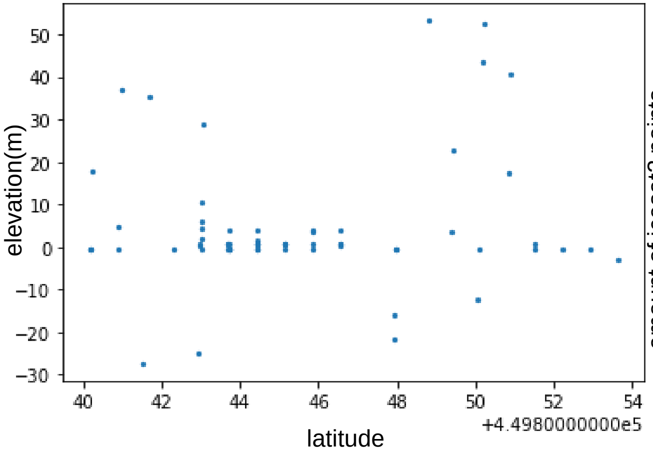
- building's footprint
- elevation and height information are used as reference

TU Delft

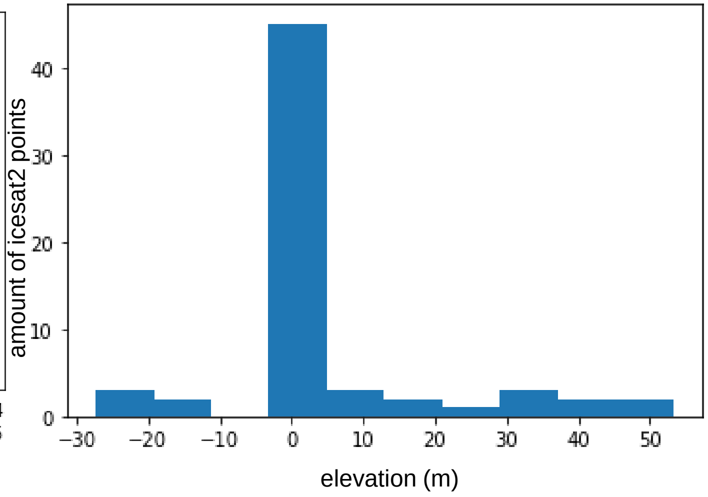
How can we get height information for each intersected footprint?



Distribution of icesat2 points



Bar plot of icesat2 points' elevation



A building footprint in Rijswijk

gid: 26367721
ground elevation: 0.377m
roof elevation_70p: 5.943m
number of intersected icesat2 points: 63

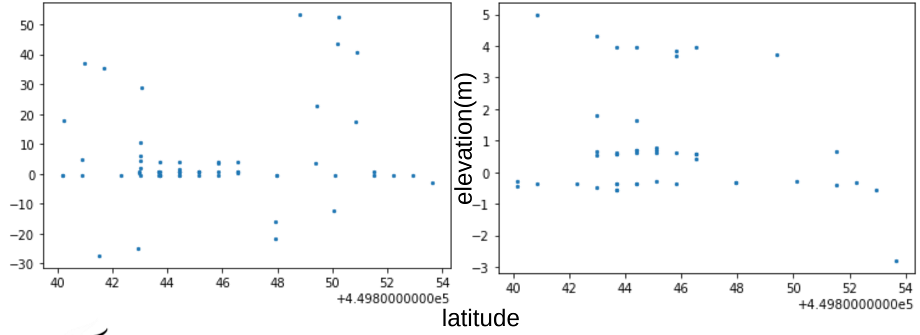
Data cleaning

ICESat-2 ATL03 product has its own noise recognize algorithm, every photon has an attribute named **"confidence"** before release.
All points in "other" category are removed.

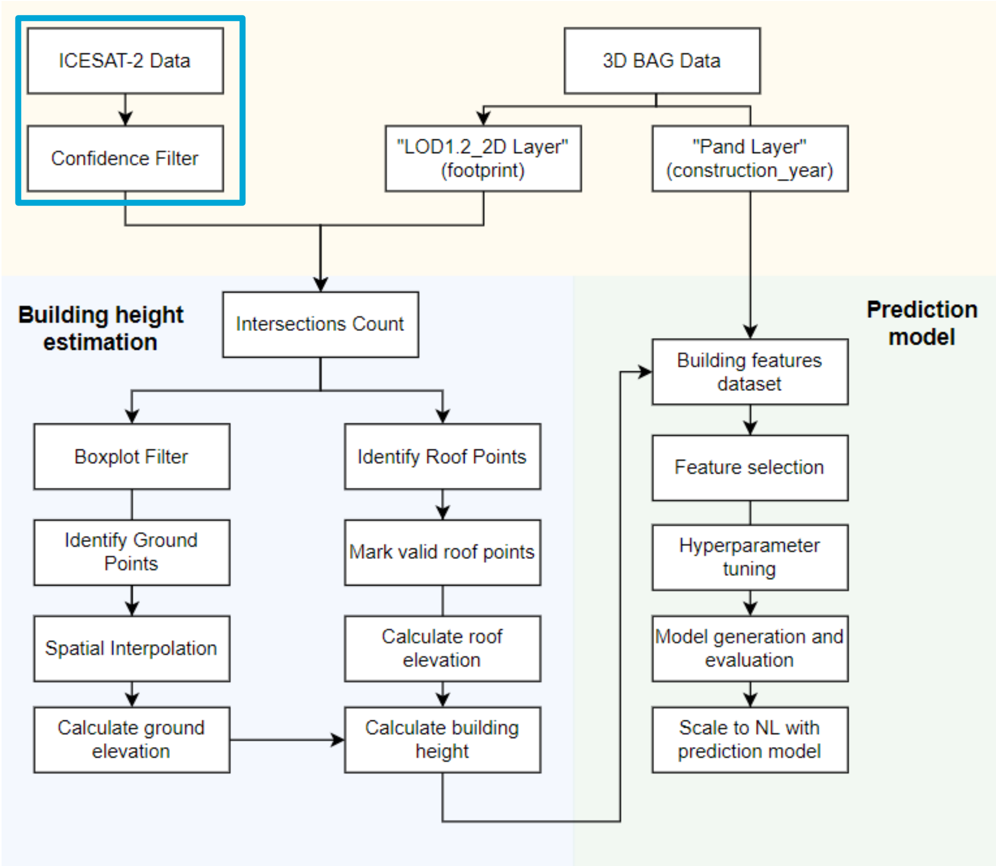
Table 3.1.: Meaning of difference confidence number

	Confidence	Description
other	-1, 0	noise
other	1	background
	2	low
	3	medium
	4	high

Distribution of icesat2 points before and after confidence filter



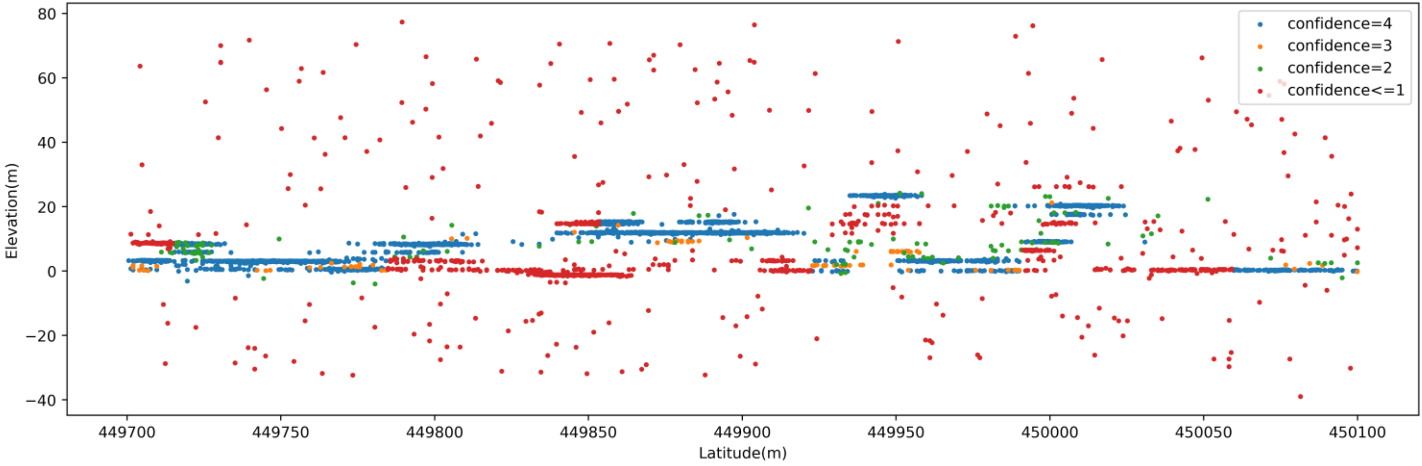
ground elevation: 0.377m
roof elevation_50p: 5.943m
number of intersected icesat2 points: 63 → 46



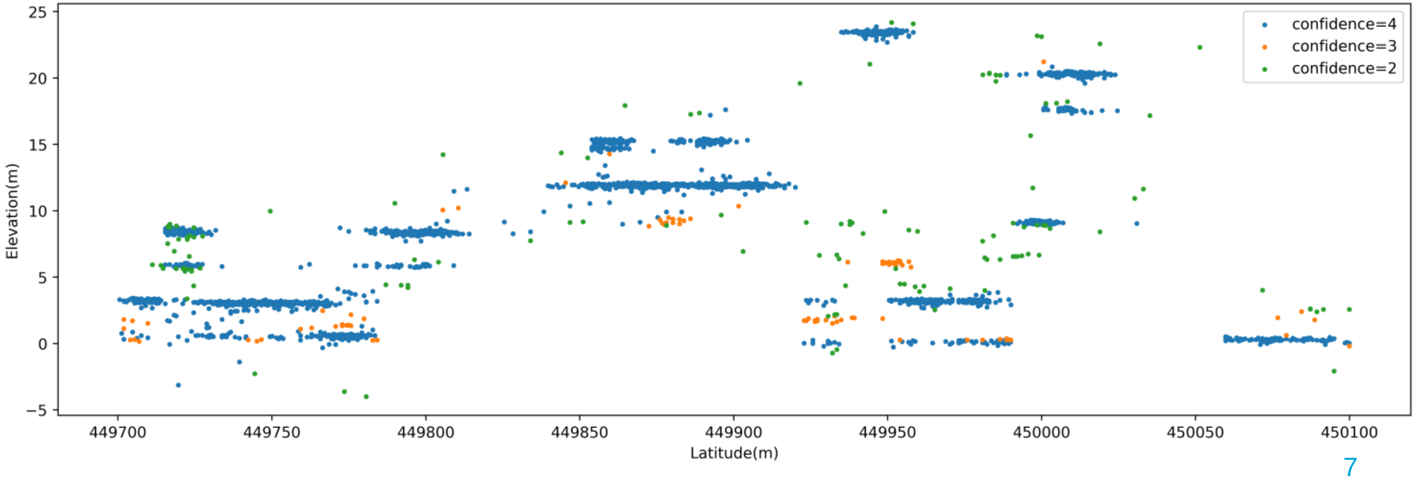
TU Delft

Confidence filter

Original data



After confidence filter



TU Delft

Data cleaning

Suppose icesat2 also has the normal distribution.

There are five important numbers in box plot: "minimum", first quartile (Q1), median, third quartile (Q3), and "maximum". The data outside the upper and lower edges is an outlier

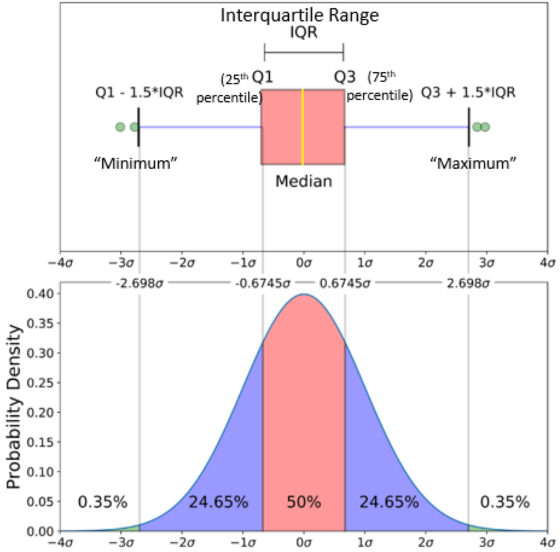
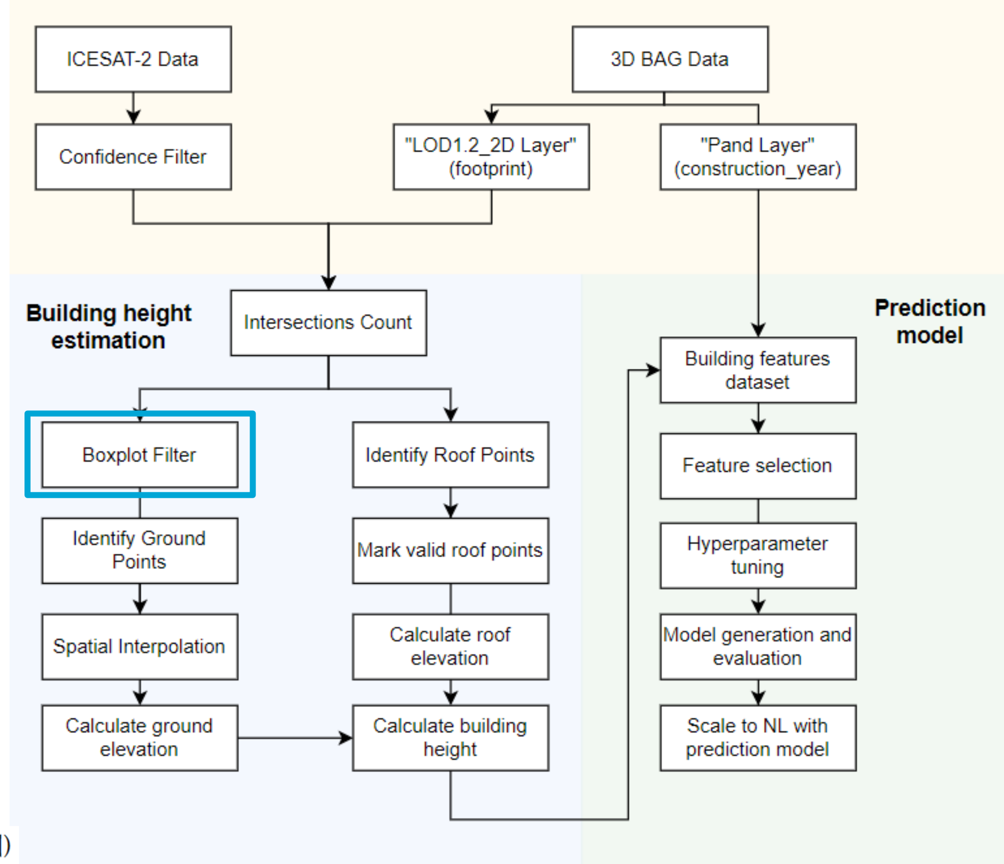
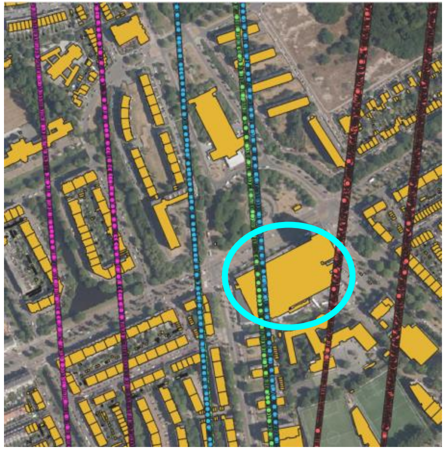
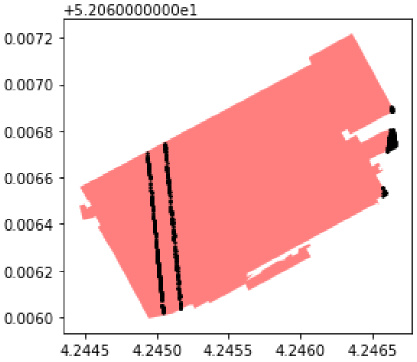


Figure 3.2.: Comparison of boxplot and normal distribution (Source: Galarnyk [18])

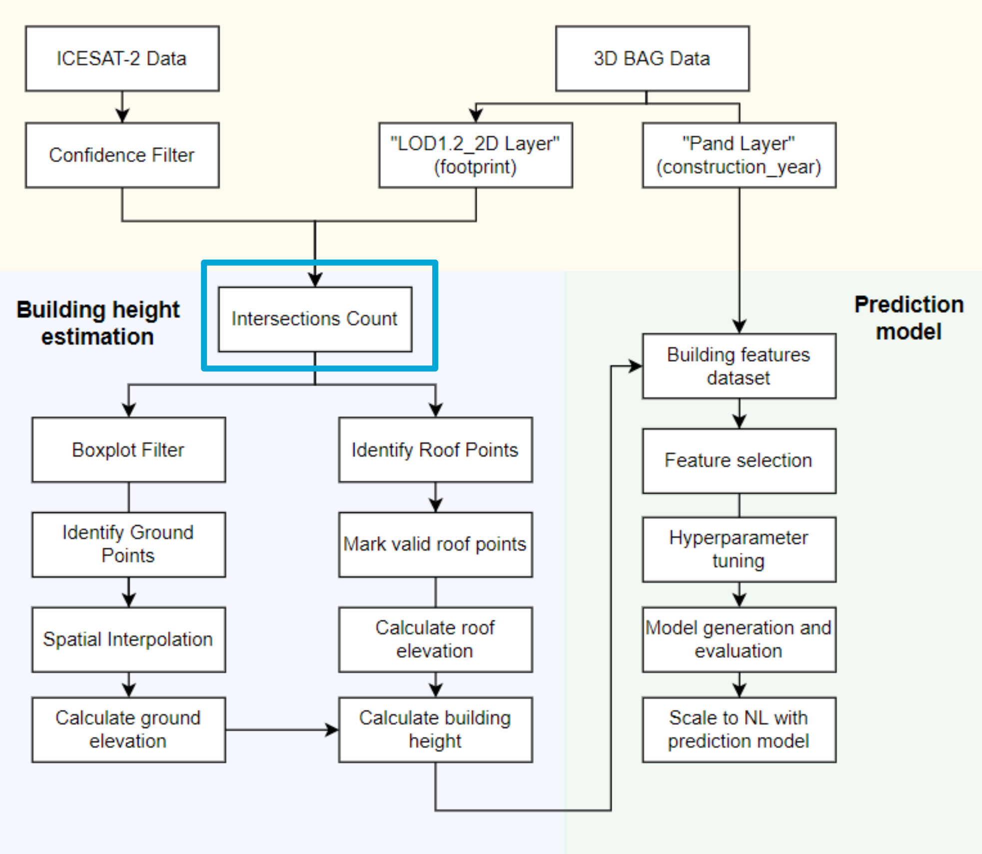


Intersection analysis





- The number of points
- The distribution of z value of these points



TU Delft

Intersection analysis

The first, most ideal type (Figure 3.3 a), perfectly contains ground data points and roof data points, which are distributed to form two distinct categories. In this case, the building height is the difference between the two categories.

The second type, containing noise (Figure 3.3 b).
It may be trees or other objects in the range. In this case, the access to ground elevation or roof elevation will be affected to some extent, and there are more challenges, such as how to distinguish between roof height and tree top height. It's hard to say which points are belong to roof.

The third type, missing data.
Because of some unknown reason, the points dropped within the footprint cannot represent the height of the ground (Figure 3.3 c) or the roof (Figure 3.3 d), or in the worst case, neither. They just missing the elevation information.
Figure 3.3 e and Figure 3.3 f illustrates that sometimes there are not enough points in the footprint to determine the ground and roof elevations of this building. There is no way to get the building height from just one or two points.

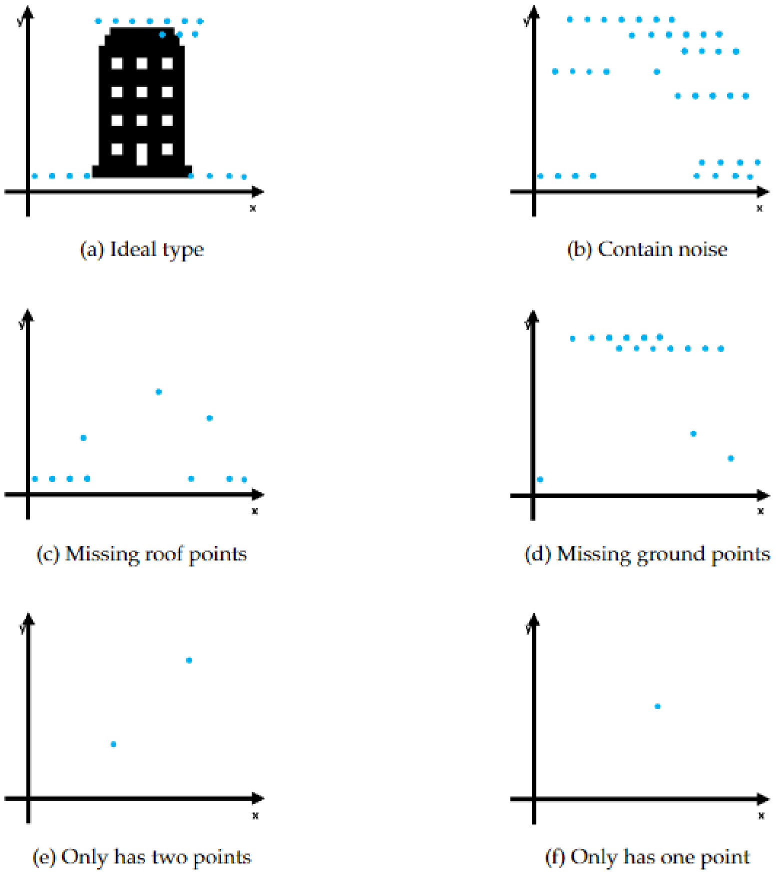


Figure 3.3.: Distribution of points in the footprint

TU Delft

Ground elevation

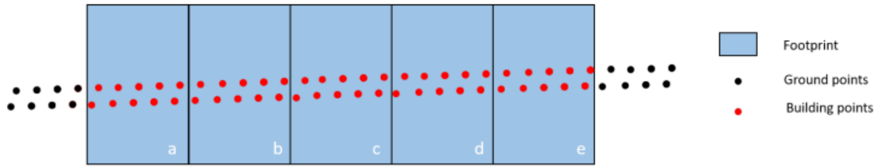


Figure 3.4.: Townhouse

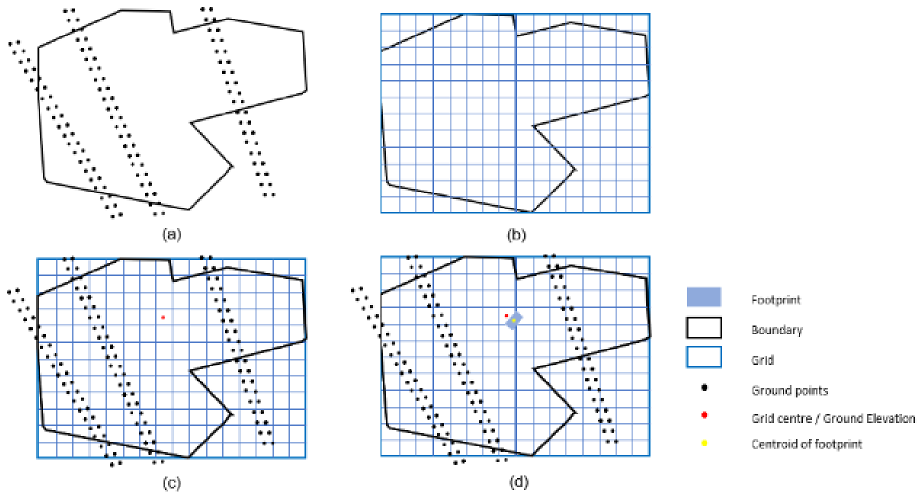


Figure 3.5.: Illustration of spatial interpolation

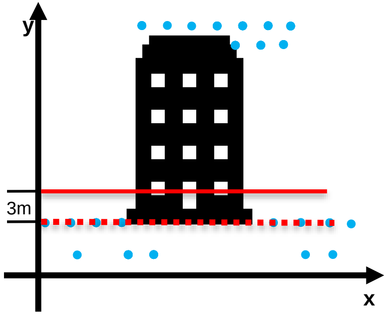
Figure 3.4 shows a situation where there is no way to obtain the ground points of a house in a townhouse complex, such as building b, building c, building d.

Grid spatial interpolation method

1. Obtain all ground points
2. Generate bounding box and grid
3. Estimate ground elevation of each grid cell's centroid
4. The ground elevation of a footprint is decided by it's nearest grid cell's centroid (red point)

TU Delft

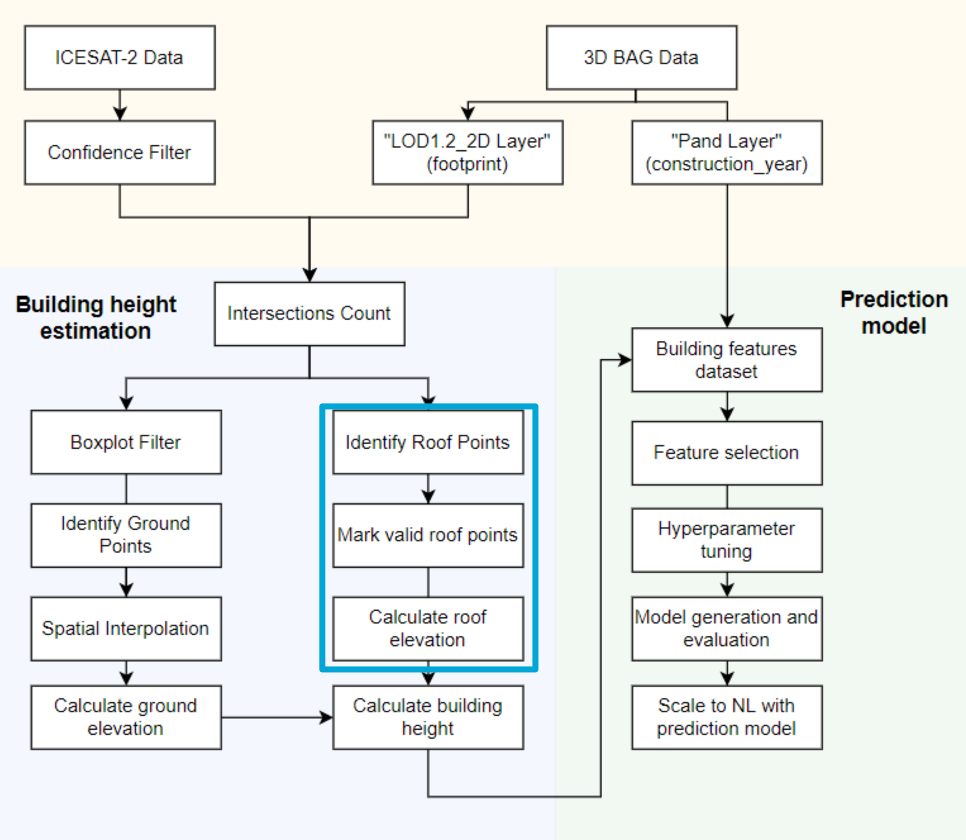
Roof elevation



Exclude those footprints that don't have points higher than the calculated ground elevation.

The minimum floor height should also under consideration. From bouwbesluit (The Dutch Building Decree), this value is should be 3 meters.

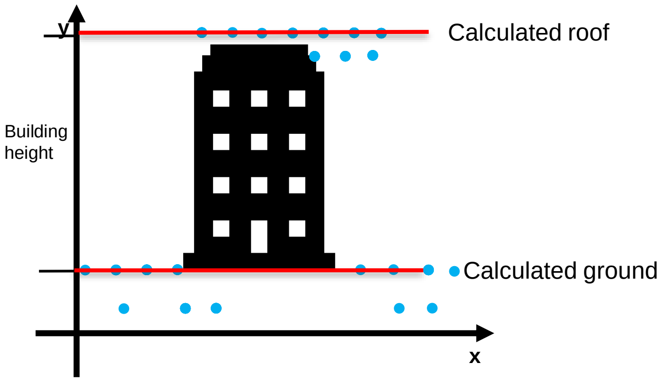
To be consistent with BAG data 50th percentile of roof points is used to define roof elevation.

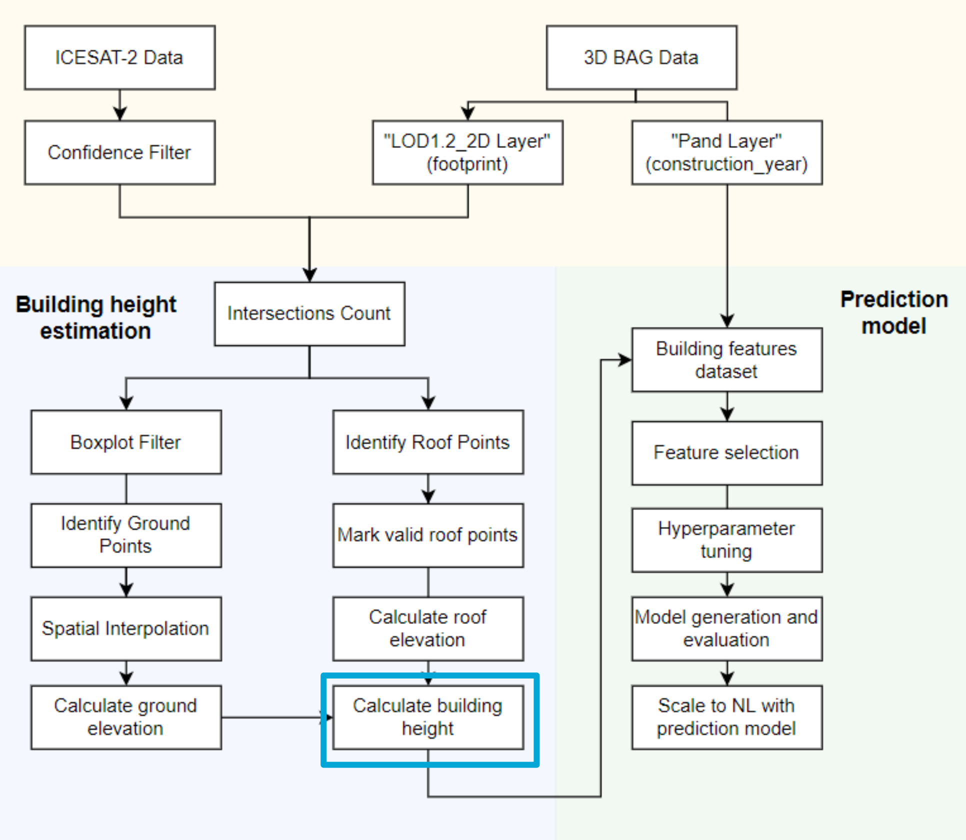


TU Delft

Building height

$$Building\_Height = Elev\_r - Elev\_g$$





Feature selection

	Feature	Description	Computation
1.	Area	The area of the building footprint	-
2.	Compactness	The Normalised Perimeter Index (NPI)	$\frac{2\sqrt{\pi A}}{P}$
3.	Number of neighbours	Buildings within a range of 100 metres of the footprint	Centroid distance
4.	Complexity	The irregularities in the footprint	$\frac{P}{\sqrt[4]{A}}$
5.	Number of adjacent buildings	Buildings within 1 metre of the footprint	Buffer intersection
6.	Length	Longest edge of MBR	-
7.	Width	Shortest edge of MBR	-
8.	Slimness	Ratio of the sides	$\frac{F_{length}}{F_{width}}$
9.	Number of vertices	Total number of vertices in the footprint	-

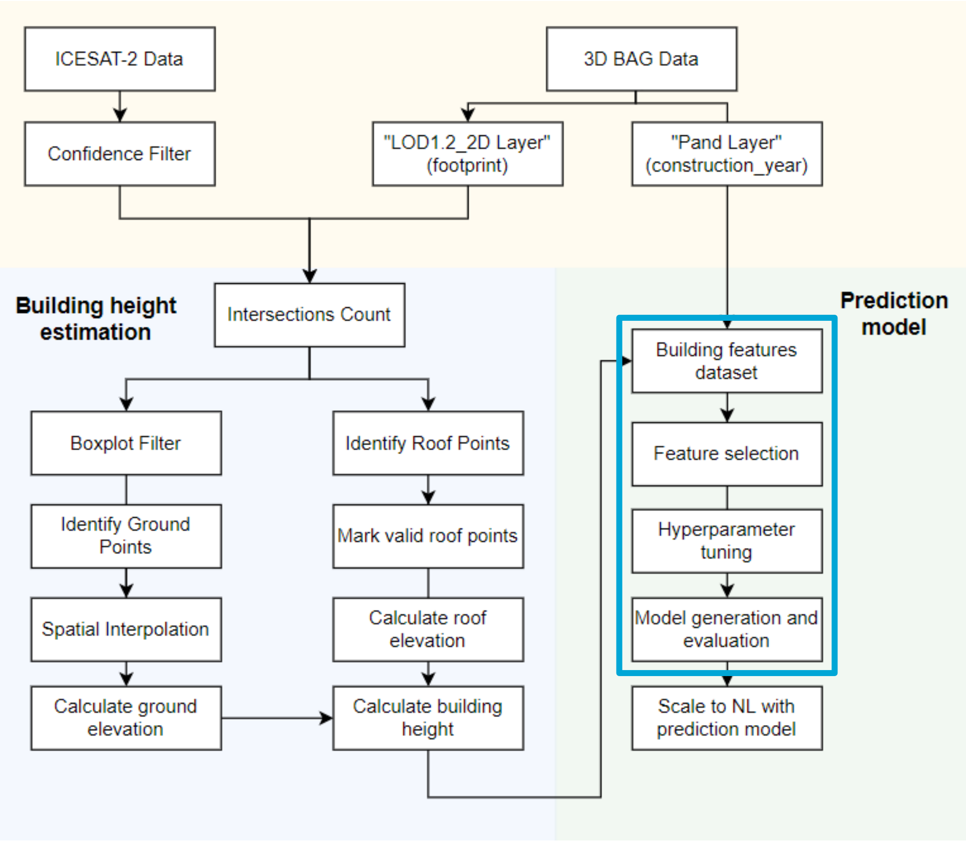
Figure 3.7.: Geometric features. *Source:*Lánský [27]

<u>plus “construction_year”, “perimeter” --- 11 features</u>

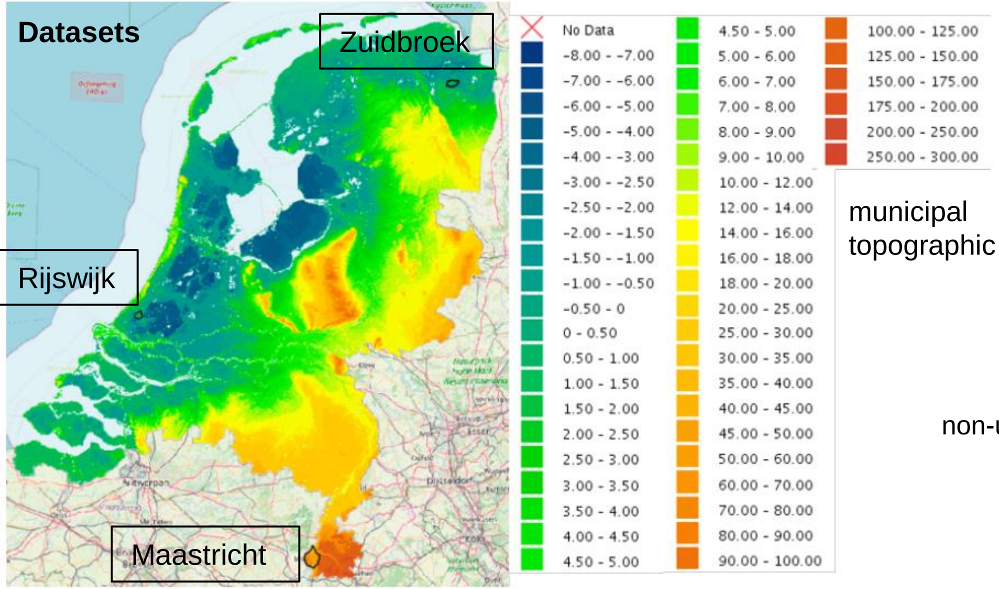
Non-geometric features are not considered, because of data availability.

- Base model
- Filter model
- Embedded model
- Wrapper model

Filter, embedded and wrapper method are from scikit-learn library

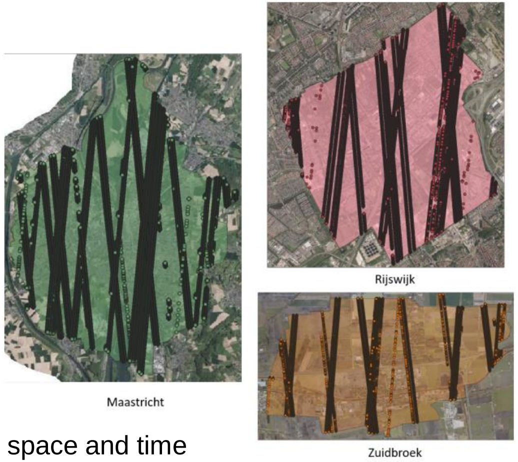


Datasets



municipal
topographic

Figure 4.2.: Location of three datasets (in black edge)



non-uniformity in space and time

Figure 4.4.: Space distribution of ICESat-2 data in three datasets

Table 4.1.: Basic information of Datasets

Municipality	No. ICESat-2 points	No. footprints	Area (km^2)
Maastricht	1,219,131	59,338	60.12
Rijswijk	597,636	17,684	14.49
Zuidbroek	146,693	2,825	17.28

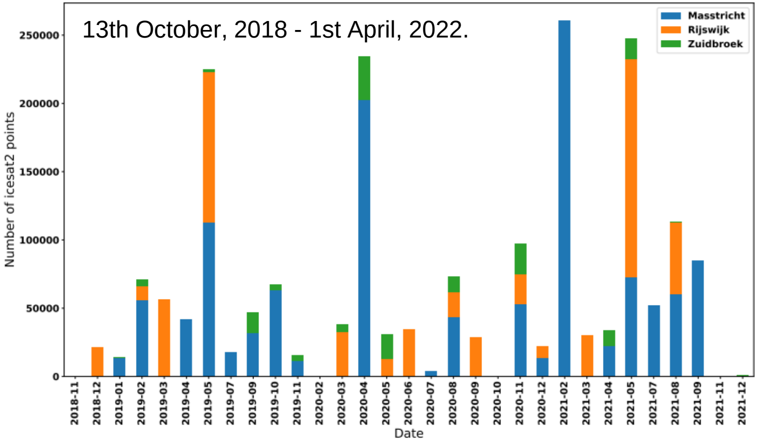


Figure 4.3.: Time distribution of ICESat-2 data in three datasets

TU Delft

Calculated building height result – Hight distribution

Table 4.2.: The amount of final valid footprints

Municipality	No. footprints	No. intersected footprints	No. intersected footprints after filter	No. final valid footprints	Final valid footprint percentage
Masstricht	59,338	2,902	2,428	1,640	2.76%
Rijswijk	17,684	1,382	723	525	2.97%
Zuidbroek	2,725	140	107	73	2.68%

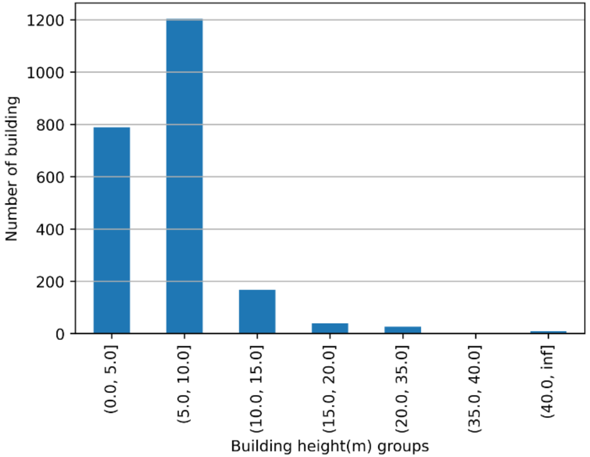


Figure 5.10.: Distribution of building height of three data sets

(0.0, 5.0]	789
(5.0, 10.0]	1204
(10.0, 15.0]	167
(15.0, 20.0]	40
(20.0, 35.0]	26
(35.0, 40.0]	2
(40.0, inf]	10

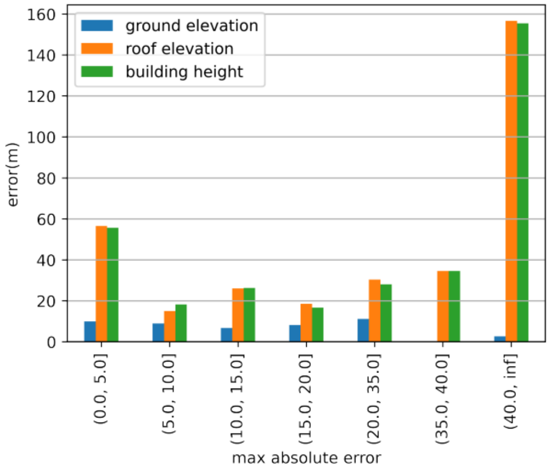
in total: 2238

There are 2238 buildings in total (results for all three data sets)

- 1204 of them have a height between 5 - 10m, more than 50%.
- And about 40% of these buildings have height in the range of 0 - 5m, which means buildings lower than ten meters tall accounted for 90% of the total buildings.
- Buildings over ten meters occupied only 10% of the total building.

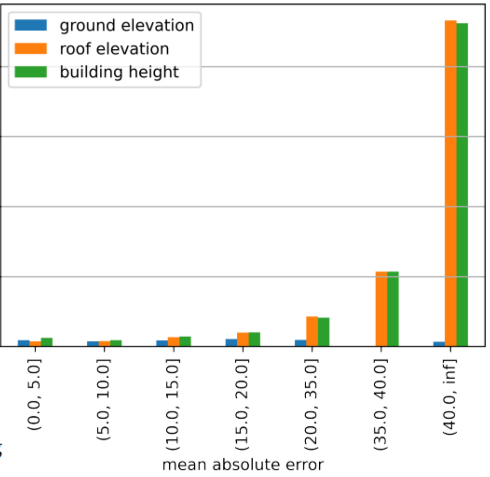
TUDelft

Calculated building height result – Error statistics



Conclusion:

- The 5-10m building height group has the largest number of buildings and the smallest mean absolute error (1.8415m).
- Then, as the building height increase, the number of buildings in each group decreases while the mean absolute error rises.
- When building height is lower than 5m (0-5m group), it has the second smallest mean absolute error (2.4752m) also the second-largest maximum absolute error (55.7198m).



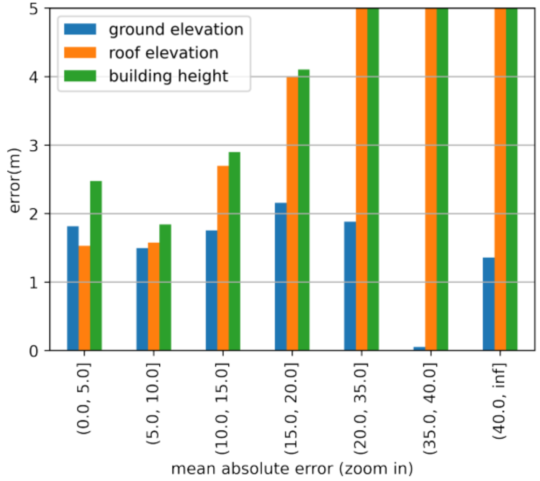
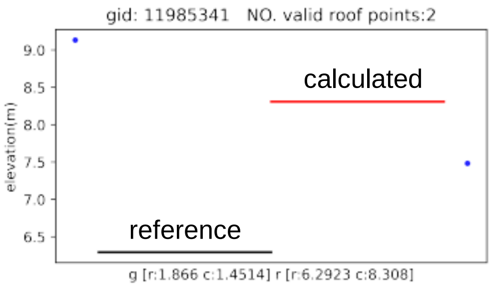


Figure 5.13.: Absolute difference of calculated and reference value among different building height levels

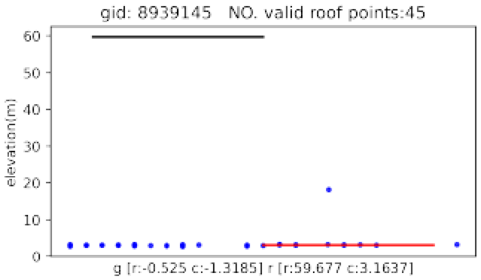
Calculated building height result –
Error cases

Each scatter plot not only shows the elevation information of ICESat-2 data inside each footprint, but also the **reference roof elevation (black lines)** and **calculated roof elevation (red lines)** in each footprint.



1) Not enough valid roof elevation data.

Two cases exist in this category.
One is that the overall number of valid roof points is tiny, maybe only one or two, which is not enough to calculate the roof elevation accurately.



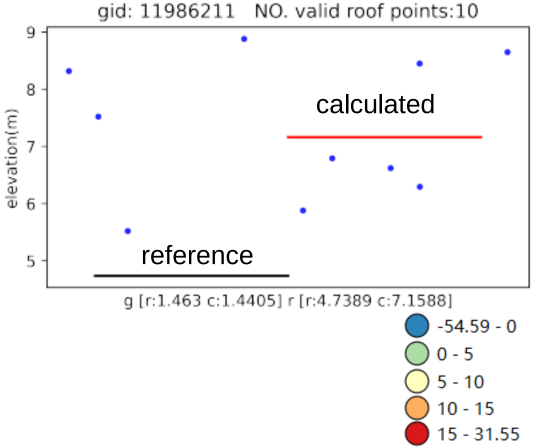
The other is that although the number of valid roof points is large, most of them do not capture the roof elevation information accurately, resulting in errors. Normally, making the calculated value lower.

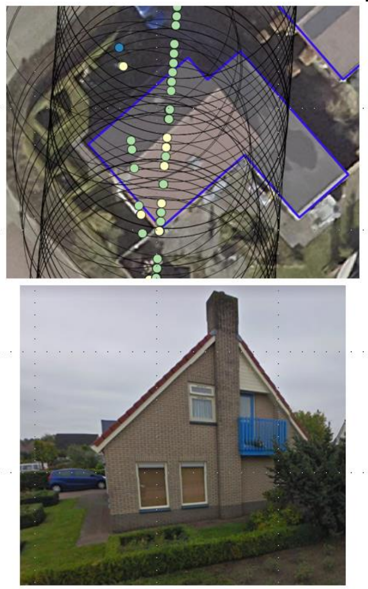
reference roof elevation (black lines)
calculated roof elevation (red lines)

Calculated building height result – Error cases

2) Irregular roof shape.

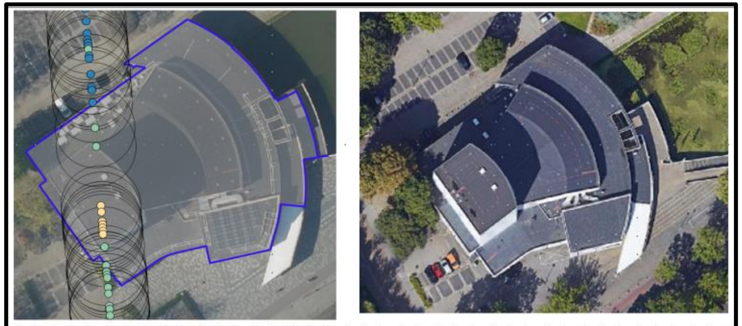
Usually, this is the case where a footprint actually comprises several parts of the roof and the elevations of these roofs are not the same.
The non-uniform roof heights introduce errors into the calculations, making the calculated value either higher or lower



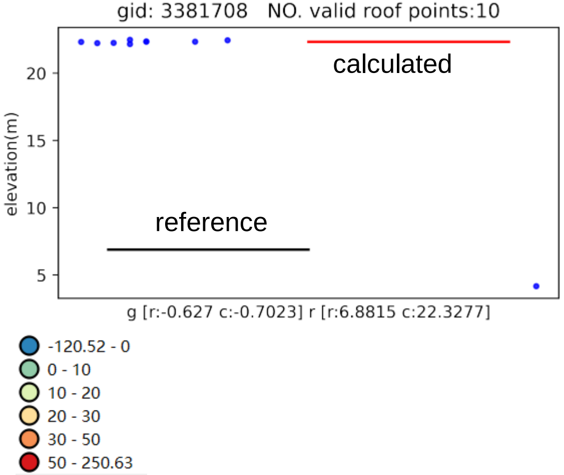


forth: gid = 11986211

a performing arts theater in Rijswijk.



Seventh: gid = 3381708



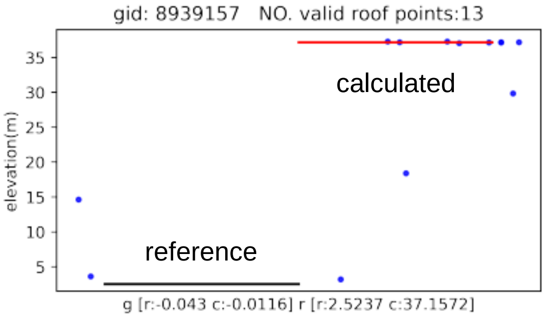
TU Delft

Calculated building height result – Error cases

3) Influence from surrounding objects.

In most cases, those footprints which are easily be influenced by surrounding objects are not the building in the traditional sense. For example, a self-built carport in the backyard, a detached garage next to a residence, and parking garage. These buildings also have their own footprint in the BAG. These buildings are lower than other buildings around them and are often in areas with high building density. Therefore, the roof elevation is easily affected by the surrounding buildings or trees because of shading.

Even though the ICESat-2 points are located inside its footprint, it may contain the elevation information of other objects. This makes the calculated value either high or lower.



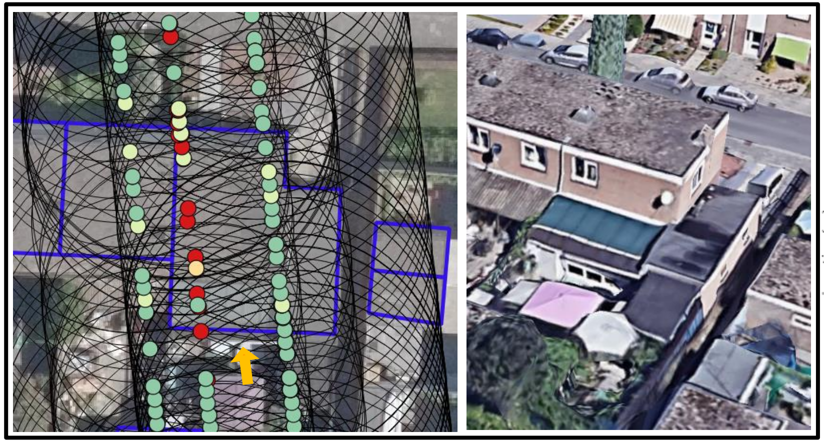
TU Delft

Calculated building height result – Error cases

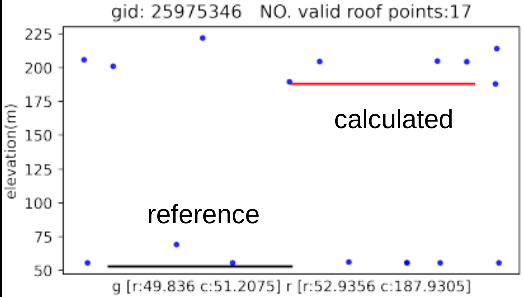
4) Effect of significant outliers.

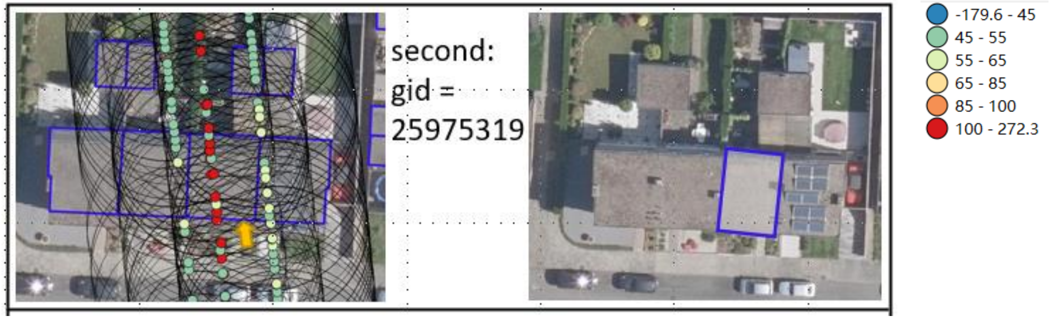
This situation is only found in the Maastricht dataset.

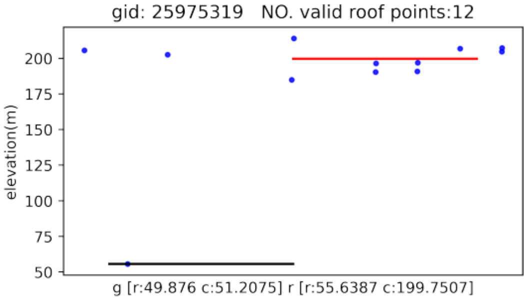
In some footprint, there are significant outliers (elevation is above 200m), causing the error.



fifth: gid = 25975346

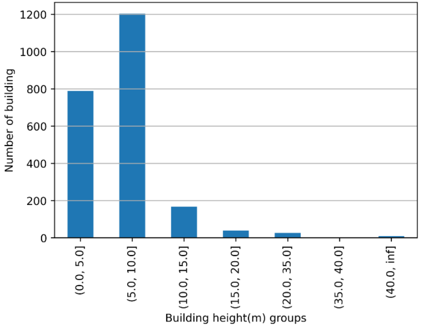






TU Delft

ML result – Model performance




```
(0.0, 5.0]       789
(5.0, 10.0]     1204
(10.0, 15.0]     167
(15.0, 20.0]      40
(20.0, 35.0]      26
(35.0, 40.0]       2
(40.0, inf]       10
```

2238 in total

Figure 5.10.: Distribution of building height of three data sets

Table 5.3.: Feature of each model

Features \ Model	Base model	Filter model	Embedded model	Wrapper model
area	✓	✓	✓	✓
perimeter	✓	✓	-	-
construction_year	✓	✓	✓	-
length	✓	✓	✓	✓
width	✓	✓	-	-
complexity	✓	✓	-	-
vertices	✓	✓	-	-
neighbour	✓	✓	-	-
slimness	✓	✓	✓	✓
adjacent_buildings	✓	-	-	-
compactness	✓	✓	✓	✓

Evaluate by metrics:

	MAE(m)	MAPE(%)	RMSE(m)	R^2	Max. error(m)
Base model	2.1305	36.6886	3.3989	0.1638	27.0906
Filter method	2.1561	36.9020	3.3967	0.1649	26.9635
Embedded method	2.2042	37.5127	3.4544	0.1363	26.9455
Wrapper method	2.2240	38.3844	3.4554	0.1358	26.9333

The data above is used to generate ML model (RFR).

If the accuracy of the model is acceptable, then it can be used to scale up to the whole of the Netherlands.

TU Delft

ML result – Model performance

Evaluate by density plot:

1) The prediction model doesn't work when building higher than 10m.

The maximum true value is 40m. However, the range of predicted values is between around 5 - 15m. This is caused by lacking data. 90% of training data is between 0-10m.

2) There is a gap between the left end of the predicted value and the true value.

This means the height of the building under five meters cannot be predicted well by all four models. This can be explained by this height group owns the second-largest maximum absolute error.

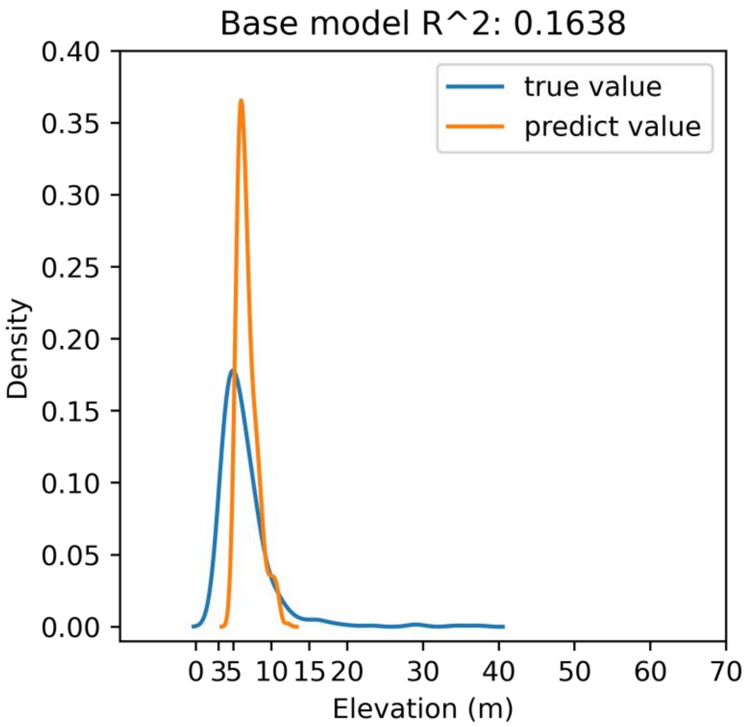


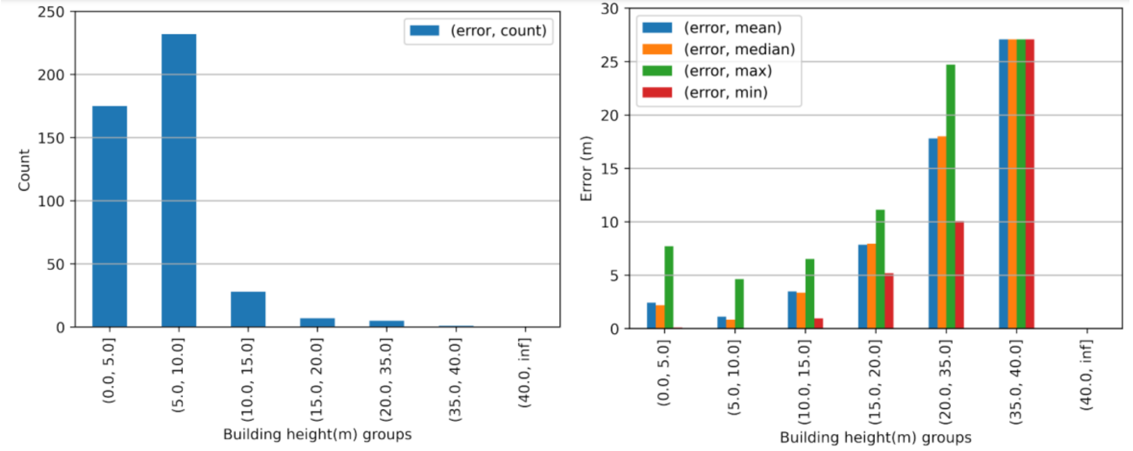
Figure 5.22.: The density plot of reference and predict value

ML result – Model performance in test dataset

Test dataset:

About 90% of the buildings are between 0-10 meters in height and only 10% of the buildings are above 10 meters. But there are no buildings with heights greater than forty meters.

It can be seen the building in 5-10m group has the smallest mean error (1.1267m).
The 0-5m group is the next one with a mean error of 2.4243m.
Then, the error gradually increases with the increase in building height. The same pattern of variation was observed for other metrics (median, maximum, and minimum error).



(a) Building height distribution

(b) Error distribution

true	count	mean	median	max	min
(0.0, 5.0]	175	2.424301	2.198502	7.716256	0.105224
(5.0, 10.0]	232	1.126729	0.853404	4.631425	0.004302
(10.0, 15.0]	28	3.488452	3.371841	6.516312	0.957918
(15.0, 20.0]	7	7.856573	7.965960	11.142793	5.187271
(20.0, 35.0]	5	17.811923	18.006745	24.722529	10.090686
(35.0, 40.0]	1	27.090629	27.090629	27.090629	27.090629
(40.0, inf]	0	NaN	NaN	NaN	NaN

Figure 5.22.: Building height and error distribution in test data set

Conclusion

Q: Can the height of all buildings in the Netherlands be estimated from ICESat-2 data and what accuracy can be achieved?

A: It is impossible to get the height of **all** buildings in the Netherlands with ICESat-2 data. But it is a feasible option for buildings between 5 and 10 meters in height.

Two main reasons:

a) few available data(quantity)

1. Sparsity and non-uniformity of the ICESat-2 data.

2. Requirements from building height estimation further reducing the amount of valid ICESat-2 data.

After filter out footprints without valid ground and roof points, this amount was reduced to less than 3%.

3. Lack of data for buildings over ten meters (only 10%).

Buildings under ten meters accounted for 90% of the total data used in model generation. This results in not enough training data for buildings over ten meters. It makes the final generated model perform poorly overall.
However, it still obtained an acceptable performance in (5,10]m height group. The MAE is 1.1267m.

b) accuracy of ICESat-2(quality)

4. The problem of data precision of ICESat-2 data.

The elevation of a ICESat-2 point falling in a footprint is influenced by other surrounding objects.

That is, even if a point falls in a footprint, it still could not provide the accurate elevation information of that footprint. Each photon point from ICESat-2 has a footprint of approximately 17m in diameter. In theory, the elevation obtained by ICESat-2 point could be any object inside this diameter.

TU Delft

What's next:

Add more training data

Improve prediction model performance when building higher than 10m

Scale to NL

Collect all building features in NL, estimate their building height with prediction model.

Thank you!

Questions?

TUDelft