Assignment 4, CIE4604, 2021-2022, SimVis: Point Cloud Classification

Teaching Assistant: Mieke Kuschnerus: m.kuschnerus@tudelft.nl

Lecturer: Roderik Lindenbergh: r.c.lindenbergh@tudelft.nl

- *Make a short report of your assignment answers in pdf format and upload it to BrightSpace until Tuesday, December 21, 10:00 o'clock. Describe your results s.t. they are reproducible. Always motivate your answer. Alternatively, submit a Jupyter Notebook, but make sure you motive your answers.*
- *Put your name and student number on the assignment.*
- *Use predefined* `Python` *commands or your own program. Python package* `sklearn` *is recommended as it supports Random Forest and neighbor search. It is possible to use alternative software, like* `R` *or* `Matlab`*, but document your choice.*

Goal of this assignment is to classify AHN-4 point cloud data using Random Forest. This description consists of two parts: (i) preparation, and (ii), the actual assignment.

Preparation. A test data set of Noordwijk aan Zee is available at:

`https://surfdrive.surf.nl/files/index.php/s/bOKKhG7JgjXbkRk`

The provided data set has been enriched by RGB colors. Random Forest classification needs features that characterize a location. In this assignment you will consider two types of features: observed and hand-crafted. Observed features are raw measurements, like signal strength, while hand-crafted features are composed values computed by you, for example the difference in signal strength between a given point and its 10 nearest neighbors.

A good tool to visualize and partly process point clouds is the Open Source *CloudCompare* software, available from http://www.danielgm.net/cc/

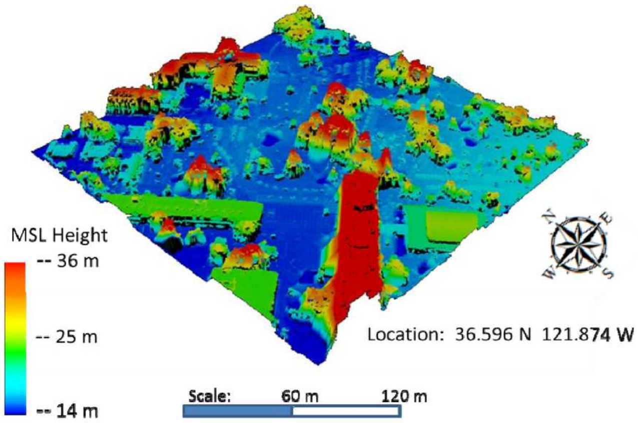

Figure 1: Point cloud visualization

a) Investigate the data set: describe each of its columns and assess the spread in values in each column. What is the meaning of each attribute?

b) What are ways to create a subset? Select a suitable subset of the data and visualize it in 3D. Start small, and if computer and software permits, try a bit larger subset.

c) Identify at least four different object classes. Choose your classes such that together these cover a large majority of your points.

d) Analyse which spatial scales and spectral properties are useful to distinguish your classes.

e) Extract training data for each of your object classes. Divide your training data in two parts, one for training, and one for validation. What could be the influence of imbalances in your training data? How could your division in training and validation data affect your results?

f) Find a suitable implementation of the Random Forest algorithm. What are its parameters? What would be good settings for these parameters, given your classification task? Apply Random Forest on your data using only your observed features, to make sure your setup is correct.

Assignment Questions

1. Describe the properties of the data (sub)set that you will classify: (i) How many points? (ii) What area does it cover? (iii) What observed features will you use? (iv) Visualize your final subset, including the useful observed features. (vi) Describe your at least four different object classes;

For each point determine neighborhoods of $k_1, k_2, ...$ points (using e.g. an efficient data structure), and use these $k_1, k_2, ...$ points to estimate several feature values.

2. Describe $\geq$ 20 different geometric attributes, obtained using at least 2 different neighborhood sizes, to characterize your points. What are the dimensions of the data covariance matrix you use to compute the geometric features? Give one example on how you determine the PCA eigenvalues of one k-neighborhood. Indicate for each feature how it could help to distinguish your classes, given also the neighborhood sizes you consider.

3. Compute all geometric features for all points in your subset using Python. Visualize selected results, e.g. by combining features in a false color visualization and/or using histograms. Which features are best at discriminating your classes? Why?

4. Describe and visualize your training data. Is your training data balanced? Make sure that a zone of your point cloud data is really ’unseen’, that is that no training data is taken from that zone, so you can inspect if Random Forest also works there.

5. Feed your training data to Random Forest using at least 10 of your best geometric and observed features. What settings did you use?

6. Classify your point cloud data and visualize and discuss your result. What went well? Give also examples where the classifier mixed up classes. What are possible explanations for these confusions? How are the classification results on the unseen zone?