**CIND 820**

Project Abstract

Student Name:

Ryan Boyd

Student Number:

501072988

Date Submitted:

1/24/2022

Submitted to:

Instructor: Tamer Abdou

Project Supervisor: TBD

Ryerson University, Chang School of Continuing Education

Project Theme:

Spatial clustering of LiDAR LAS point cloud datasets using Python, identifying objects, comparing and classifying points

**Background**

LiDAR data collected by aerial aircraft is freely available by many municipalities and organizations. It can be used and continues to be valuable for generating models of the earth’s surface among other things. The LiDAR datasets, often known as “point clouds”, can contain millions of data points collected from laser pulse returns which are reflected off the earths surface, giving the exact elevation of each pulse return based on the time it takes for the laser pulse to return to the sensor. LAS point cloud data can be read with python and analyzed as a NumPy array similar to any other dataset. The distinguishing feature of the data is the coordinate attributes represent a location on the earth’s surface. This could enable the potential for clustering based on spatial proximity of these points, rather than only clustering using attribute values. I would like to examine the possibility of clustering LiDAR points to identify objects and/or land cover type and to examine the feasibility and effectiveness of this task using Python libraries.

**Problem/Research Question**

These datasets include millions of data points across a certain region of the earth, giving fine grained information about the surface. The data points however are independent of one another in their raw form. Grouping them together and classifying them is not performed out of the box and needs to be done independently after collection of the data.

The data points from lasers that hit the ground can be used to create a model of the surface of the earth. Many points however strike other things that are located above the ground. Using clustering algorithms on these points might be useful in attempting to identify and distinguish these objects.

I will attempt to answer the following research questions. What are some practical use cases of spatial clustering algorithms on lidar point cloud data? Can spatial clustering algorithms be used to detect objects in LiDAR point clouds? How do the available algorithms perform and are they effective given the appropriate parameters? What parameters are significant and how might we improve the performance of the models?

**Proposed Solution**

I will use spatial clustering algorithms to attempt to locate/detect objects using lidar points based on their similar characteristics and spatial proximity. Since physical objects will be contained within a certain localized area and also share similar characteristics based on the materials they consist of, I will investigate whether clustering algorithms are able to identify individual objects such as cars, buildings, tree canopies or potentially even lengths of roadways or bodies of water etc.

I expect that there will be certain characteristics that might be used to identify objects in addition to their proximity. Certain surfaces may return multiple laser points for example, which suggest the surface is partially penetrable, some surfaces reflect more light and therefore return a very strong intensity while other surfaces absorb more light and return a lower intensity pulse, these characteristics could offer clues to what the surface might be i.e. vegetation, metal or water. Although I don’t expect to classify all the surfaces, these attributes may be useful in the clustering algorithms to group similar surfaces together as an object for example. I plan to implement clustering techniques and adjust the process as needed.

Python will be used to conduct this project and some of the libraries being considered include:

* Laspy: for reading las files and converting into numpy arrays
* Numpy: for storing point data read from raw las data files and performing operations
* scikit-learn: includes the algorithms used for clustering such as DBSCAN, HDBSCAN\*, K-MEANS, KNN
* pandas/geopandas: for working with and visualizing data and spatial data in particular
* matplotlib/seaborn: for possible improved visualizations of data and results

*Tentative Model Inputs*

|  |  |  |
| --- | --- | --- |
| Attribute Name | Description | Data Type |
| X (longitude in meters) | Location on horizontal axis | Double |
| Y (latitude in meters) | Location of vertical axis | Double |
| Z (elevation in meters) | Height from the ground | Double |
| Intensity | “Brightness” of the laser pulse returned to the sensor | Integer |
| Total number of points returned | Number of separate pulses returned to the sensor | Short Integer |
| Return number | Return # of the given data point | Short Integer |
| R, G, B (if available) | Red, Green and Blue values of the imagery at each point | Array or 3 separate integer values |
| Classification | Type of surface or object, (i.e. building, ground, water, etc)  \*see below | Integer |

All data points can/will be normalized before being used in the given algorithms to improve performance.

*Tentative Algorithms/Methods*

DBSCAN (Density Based Spatial Clustering for Applications with Noise)

<https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html>

HDBSCAN (Hierarchical Density Based Spatial Clustering for Applications with Noise)

<https://pypi.org/project/hdbscan/>

K-Means Clustering

<https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>

kNN Clustering (possibly used for classification if needed, see below)

<https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>

**Datasets**

Raw and/or classified LAS “point cloud” datasets from government open data sites.

*Sample Data Sources*

Ontario Lidar Datasets:

<https://open.canada.ca/data/en/dataset/6a0c7177-24de-4eee-89dd-ef7abef427ff>

British Columbia Lidar Datasets:

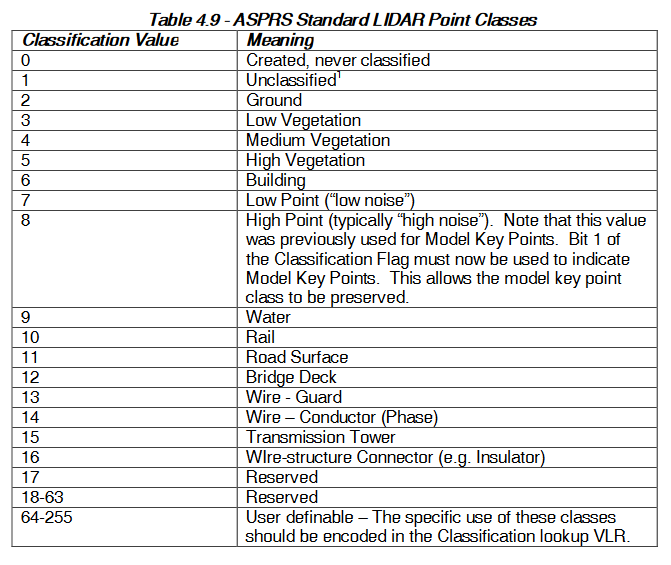
<https://governmentofbc.maps.arcgis.com/apps/MapSeries/index.html?appid=d06b37979b0c4709b7fcf2a1ed458e03>

New Brunswick Lidar Datasets:

<https://open.canada.ca/data/en/dataset/3d6fdf9d-585e-5d49-5a85-922efe568e35>

Exact dataset and area of the location is yet to be decided, but all LAS data files are similar, so this could essentially be applied to any dataset depending on the size and speed requirements.

Data points may be classified beforehand depending on the original dataset, the standard classification for LAS data is given as follows:



Source: <https://www.asprs.org/a/society/committees/lidar/LAS_1-4_R6.pdf>

For unclassified points (or never classified), an attempt will be made to classify them if feasible as part of this project. Otherwise, they will be considered unclassified for the purposes of the study and used in the clustering algorithms as such. The clustering algorithms would simply consider the class as an additional attribute to be considered when carrying out the clustering, in addition to the spatial locations of each point.

*Sources*

<https://www.asprs.org/a/society/committees/lidar/LAS_1-4_R6.pdf>

<https://www.asprs.org/wp-content/uploads/2010/12/LAS_Specification.pdf>

<https://governmentofbc.maps.arcgis.com/apps/MapSeries/index.html?appid=d06b37979b0c4709b7fcf2a1ed458e03>

<https://en.wikipedia.org/wiki/LAS_file_format>