A picture containing text

Description automatically generated

**CIND 820**

Literature Review and Data Description

Student Name: Ryan Boyd

Student Number: 501072988

Date Submitted: 2/14/2022

Project Supervisor: Dr. Ashok Bhowmick

Contents

[Introduction 3](#_Toc95762682)

[Updated Research Question 3](#_Toc95762683)

[Literature Review 4](#_Toc95762684)

[Data Format and laspy library 4](#_Toc95762685)

[Data Preprocessing 4](#_Toc95762686)

[Data segmentation or Subsampling 5](#_Toc95762687)

[Classification operations and algorithms 5](#_Toc95762688)

[Combination with Aerial Imagery 5](#_Toc95762689)

[Calculated attributes 5](#_Toc95762690)

[Buildings detection 6](#_Toc95762691)

[Trees or forest canopy detection 6](#_Toc95762692)

[Roads or Pavement and ground 6](#_Toc95762693)

[Other classes 7](#_Toc95762694)

[Object edge or plane detection 7](#_Toc95762695)

[Supervised vs Unsupervised algorithms 7](#_Toc95762696)

[Differentiation of Objects 8](#_Toc95762697)

[Descriptive Statistics 8](#_Toc95762698)

[Dataset Statistics 9](#_Toc95762699)

[Github Repository 10](#_Toc95762700)

[Proposed Methodology 11](#_Toc95762701)

[References 12](#_Toc95762702)

# Introduction

This project will seek to investigate the use of spatial clustering on LiDAR point cloud data and whether spatial clustering is able to effectively differentiate objects from the larger dataset or point cloud. In addition, classification methods will be performed to attempt to classify the points based on their similar characteristics. Python and relevant libraries will be used to attempt this.

LiDAR can be very useful for gathering information about the physical composition of an area. Millions of precise data points, usually within a few cm of one another, are collected, providing information about the elevation and composition of a specific location. By analyzing these points we can infer descriptive information about the surrounding area. [10]

There appears to be many methods for analyzing this data, some of which I will explore in this project, while some other methods are more advanced and will be out of the scope of this project. These more advanced topics however may still be explored to provide background information about the subject matter.

I believe this project is useful to explore the capabilities of remote sensing data such as lidar and aerial imagery from a data science perspective. There have been many advancements and much research work done in analyzing similar LiDAR data and I will attempt to emulate elements from these works.

# Updated Research Question

Are available open-source classification and spatial clustering workflows practical and effective for distinguishing objects in LiDAR point cloud data, how well do they perform, how might they be improved upon and what are some observed drawbacks?

# Literature Review

In my research for literature on the topic of this project, I have found that there are several advanced methods for applying classified machine learning algorithms to lidar data to classify points and identify objects. I have found many relevant articles spanning a wide range of applications offering insight into the scope of this project. While most of the work done in the articles found will not be attempted by me in this project, I believe they are important as they might provide insight into the work I do in classifying and clustering the points. I found a lot more information on the use cases of LiDAR classification algorithms, but somewhat limited work on the unsupervised clustering methods. I believe classifying the data will likely be a major first step in being able to cluster the points successfully and will therefore focus on this as well. In the review of the literature, I have discovered there may be a necessity/importance to classifying the points prior to clustering.

## Data Format and laspy library

The LAS data format is an industry standard for LiDAR data point clouds stored in binary and is specified by the American Society for Photogrammetry and Remote Sensing [14]. The laspy library for interacting with LAS files is very popular and capable. [https://pythonhosted.org/laspy/index.html, 14]

## Data Preprocessing

Data preprocessing involves cleaning or preparing the data before performing any major analysis with it or running any models with the data. These steps might include removing noise or any extreme outliers that may be erroneous data, normalizing points to better fit the models, or filtering the dataset by any requirements such as the spatial boundary of the study area. The specifics about these steps will be determined as required by the manipulation of the inputs given the results of the work. Most of the articles I found, did some sort of preprocessing steps prior to performing the bulk of the work.

## Data segmentation or Subsampling

Splitting up the large amount of data might be a necessary step to efficient processing[7,18]. To simplify the dataset without losing too much valuable information, subsampling can be completed on the data in order to reduce the number of similar points within a given area, essentially thinning it out to achieve similar results with less processing time required.

## Classification operations and algorithms

While the intended focus of this project was on the clustering operations, there are many use cases of classification algorithms with LiDAR data. Some of the classification methods include Decision Trees i.e. Random Forest [2,12], CNN [1,5], Support Vector Machine [7], RANSAC [6], among other deep learning and machine learning algorithms that have been explored [5]. Many of these algorithms are available in the sci-kit learn python package. The main issue with the classification in this case is that I may not have access to perfectly classified point data for the use of training unless I produce it myself. For this reason, I may attempt the classification and visually inspect, or create the training data manually in order to assess the performance of the classification algorithms. This would have to be done over a small sample space of the whole dataset to facilitate the manual work needed.

## Combination with Aerial Imagery

There is a strong case for using aerial imagery in conjunction with LiDAR data which is explained in many of the articles I have referenced [3,4,8,15,16]. The R,G,B values can be extracted for each point location and added as columns to the numpy array containing the las data points. I have confirmed the aerial imagery for the same year is available for the given dataset and will be attempted for this purpose.

## Calculated attributes

In addition to the provided attributes that are available with the LAS datasets, there are some that can potentially be created from the data such as the eigenvalue/eigenvectors [2,5] which can be used to group the points to a common surface that can be used to improve the clustering. This will be examined depending on the complexity of doing so but may be found to be out of the scope of this project.

## Buildings detection

Detecting the boundaries of building footprints is a major practical use case of LiDAR data since this information is very useful for a variety of reasons [10]. While there exist several techniques for doing this, such as training a neural network classifier for example, the LiDAR data proves to be very useful for this as seen in the research conducted [3,6,15,16]. Since building footprints detected by LiDAR points have certain characteristics such as height off the ground, uniform shape and surface [10] there are methods that can take advantage of these. Depending on the complexity and availability of the classifiers in question, building classification may be attempted in this project for the given data/area.

## Trees or forest canopy detection

Trees and forest are another major class of objects that can be collected from the LiDAR datasets. [4,9,10,12,13]. Typically, NIR or multi-spectrum imagery is useful for detecting vegetation [4,12]. Since I will not be using multi-spectral imagery I will rely on the classification techniques mentioned previously. Since the LiDAR data can provide multiple return pulses that allow beams to “see through” layers of vegetation before finally hitting the ground [10], we may be able to use the “Number of Return Pulses” attribute from the LAS dataset to facilitate classifying the vegetation points.

## Roads or Pavement and ground

Roads can also be detected using lidar data based on some determined conditions such as the intensity values of the points since the intensity of the light pulses returned off of asphalt/pavement tend to fall within a similar or predictable range [15]. The typical shape of roads can be used to detect where there might be roads when considering the surrounding landscape [10]. It may also be possible to identify the edges of these roads using some techniques, as well as the measuring the slope[10]. By using a combination of expected characteristics of road features, it may be possible to classify points which belong to a road in the given point cloud.

## Other classes

Many other classes or types of objects exist such as: Water, Bridges, Railways, Power Lines/Wires, Poles, Vehicles, etc., all with unique methods of detection using the data [1,20]. These classes will likely remain unclassified or may be classified as an unknown class represented only by a number.

## Object edge or plane detection

There exist several techniques to detect flat common surfaces in order to better detect the surfaces of objects. These include edge detection such as the Canny algorithm [3], calculating the eigenvalues [2] for a small subregion of points, and using the RANSAC algorithm for detecting planes (flat surfaces) in 3D data or point clouds [6,7]. With these methods we can determine where edges or surfaces exist in groups of points for a local region. This may be useful to improve the clustering of objects. This will be explored depending on the initial results and level of complexity of these tasks as the mathematics seems somewhat more advanced.

## Supervised vs Unsupervised algorithms

Due to the lack of labelled data, I will focus mainly on unsupervised methods. Wherever classification is attempted and labels are needed, classes may need to be manually created to train the given classifier. Alternatively, an unsupervised classifier such as k-means may be used. By classifying the data first, I aim to improve the spatial clustering performance by clustering only on the classified data. Essentially, the classified data will most likely be unlabelled and simply classed as a sequential number without meaning.

## Differentiation of Objects

The focus of this project will be to attempt to cluster objects together but may not be able to determine what the actual objects are. Many algorithms cited in relevant papers use proprietary methods or are outside of the scope of difficulty for this project. Many papers also were not available to be viewed for free without a license and the full details remain hidden behind a paywall. The objective of this project will be to investigate options using reasonably simple or straight forward methods to achieve the results wherever plausible. Essentially, most points will remain unclassified.

# Descriptive Statistics

The dataset has been selected based on a number of predefined criteria: having a few man-made objects such as buildings, cars and roadways, as well as natural objects such as trees, bushes and bare earth. The aerial imagery was selected for the same year. The area is approx. 90,000 square meters or 0.09 square km located just south of the Chinatown neighbourhood in Vancouver, BC.

|  |  |
| --- | --- |
| A picture containing text, fabric  Description automatically generated | An aerial view of a city  Description automatically generated |
| 2016 Lidar Point Cloud shown in 2D  Symbolized by elevation/height:  Blue = Low (0m), Red = High (100m) | Aerial Imagery 2016 |

Total Point Count (Filtered to Smaller Area): 864,440

Point Density/resolution: ~7-10 points per meter

Pre-Classified Points:

Ground (class 2): 74,748

Unclassified (class 1): 789,692

## Dataset Statistics

|  |  |
| --- | --- |
| X Min | 492042.58 m |
| X Max | 492347.95 m |
| X Range | 305.37 m |
| Y Min | 5458341.12 m |
| Y Max | 5458646.49 m |
| Y Range | 305.37 m |
| Z Min | -9.26 m |
| Z Max | 106.62 m |
| Z Range | 115.88 m |
| Intensity Range | 0-255 |
| Point Return Count Range | 1-8 |
| Z (Height) Distribution: | Point #  Chart, histogram  Description automatically generatedm |
| Intensity Distribution: | Point #  I |

# Github Repository

The following github repository has been set up and will host code and important files:

<https://github.com/rboyd-ryerson/CIND820>

# Proposed Methodology

Diagram

Description automatically generated

# References

1. *Øivind Due Trier, Jarle Hamar Reksten, Kristian Løseth.* **Automated mapping of cultural heritage in Norway from airborne lidar data using faster R-CNN**. International Journal of Applied Earth Observations and Geoinformation, Issue 95, 2021
2. *Esrine Chehata, Li Guo, Clément Mallet.* **Airborne lidar feature selection for urban classification using random forests**. Laserscanning, Paris, France. ffhal-02384719, Sep 2009
3. *LI Yong, Wu Huayi.* **Adaptive building edge detection by combining lidar data and aerial images.** The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B1. Beijing 2008
4. *M. A. Dogon-Yaro, P. Kumara, A. Abdul Rahman, G. Buyuksalih.* **Semi-automated approach for mapping urban trees from integrated aerial lidar point cloud and digital imagery datasets**. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLII-4/W1, 2016
5. *E. Özdemir, F. Remondino, A. Golkar.* **Aerial point cloud classification with deep learning and machine learning algorithms**. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLII-4/W18, 2019
6. *D.L. Bool, L.C. Mabaquiao, M.E. Tupas, J.L. Fabila.* **Automated building detection using ransac from classified lidar point cloud data**. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLII-4/W9, 2018
7. *Crawford, Peter, Wang, Jinfei.* **Effects of Aerial LiDAR Data Density on the Accuracy of Building Reconstruction.** Electronic Thesis and Dissertation Repository 5552. 2017
8. *Biswas Parajuli.* **Machine Learning Algorithms and Applications for Lidar, Images, and Unstructured Data.** Florida State University Libraries, Electronic Theses, Treatises and Dissertations. 2019
9. *João P. Pereira, Holger Weinacker, Barbara Koch, Marcos B. Schimalski.* **Scikit-image for trees local maxima detection.** Proceedings of SilviLaser 2015 - September 28-30, 2015
10. *Russell A. White, Brian C. Dietterick, Thomas Mastin and Rollin Strohman.* **Forest Roads Mapped Using LiDAR in Steep Forested Terrain**. Remote Sensing, ISSN 2072-4292, www.mdpi.com/journal/remotesensing, 15 April 2010
11. *Bruce D. Cook, Lawrence A. Corp, Ross F. Nelson, Elizabeth M. Middleton, Douglas C. Morton, Joel T. McCorkel, Jeffrey G. Masek, Kenneth J. Ranson, Vuong Ly and Paul M. Montesano.* **NASA Goddard’s LiDAR, Hyperspectral and Thermal (G-LiHT) Airborne Imager.** Remote Sensing ISSN 2072-4292, [www.mdpi.com/journal/remotesensing](http://www.mdpi.com/journal/remotesensing), 3 August 2013
12. *Julia Marrs and Wenge Ni-Meister.* **Machine Learning Techniques for Tree Species Classification Using Co-Registered LiDAR and Hyperspectral Data.** Remote Sensing, 2019, 11, 819; doi:10.3390/rs11070819 5 April 2019
13. *Ana Novo, Noelia Fariñas-Álvarez, Joaquin Martínez-Sánchez, Higinio González-Jorge and Henrique Lorenzo.* **Automatic Processing of Aerial LiDAR Data to Detect Vegetation Continuity in the Surroundings of Roads.** Remote Sensing 2020, 12, 1677; doi:10.3390/rs12101677 [www.mdpi.com/journal/remotesensing**.** 23 May 2020](http://www.mdpi.com/journal/remotesensing.%2023%20May%202020)
14. *Michal Polák, Jakub Miˇrijovský, Alba E. Hernándiz, Zdenˇek Špíšek, Radoslav Koprna and Jan F. Humplík.***Innovative UAV LiDAR Generated Point-Cloud Processing Algorithm in Python for Unsupervised Detection and Analysis of Agricultural Field-Plots.** Remote Sens. 2021, 13, 3169. https://doi.org/10.3390/rs13163169 <https://www.mdpi.com/journal/remotesensing>, 10 August 2021
15. *Ming-Jer Huang, Shiahn-Wern Shyue, Liang-Hwei Lee, and Chih-Chung Kao.* **A Knowledge-based Approach to Urban Feature Classification Using Aerial Imagery with Lidar Data.** Photogrammetric Engineering & Remote Sensing Vol. 74, No. 12, pp. 1473–1485, December 2008
16. *Xuelian Meng, Nate Currit, Le Wang, and Xiaojun Yang.* **Detect Residential Buildings from Lidar and Aerial Photographs through Object-Oriented Land-Use Classification.** Photogrammetric Engineering & Remote Sensing Vol. 78, No. 1, pp. 35–44, January 2012
17. *Shane Grigsby.* **Spatial Clustering in Python: Processing LiDAR points into objects.** *University of California, Santa Barbara, NSERC*
18. *Florent Poux.* **How to automate LiDAR point cloud sub-sampling with Python.** Towards Data Science. November 2020
19. Florent Poux. [**Discover 3D Point Cloud Processing with Python.** Towards Data Science](https://towardsdatascience.com/discover-3d-point-cloud-processing-with-python-6112d9ee38e7). April 2020
20. *Alex Simkiv.* **Point Cloud Data: Simple Approach; LIDAR data for power line detection.** Towards Data Science. June 2019