1. **Problem definition and description**

The main purpose of the project is to reduce the dimension of the data from 3(longitude, latitude, and altitude) to 2, and discuss its geographical meaning. To reduce the dimension, principal component analysis (PCA) was conducted.

PCA is a statistical instrument able to identify the variables explaining most variation within a sample. Therefore, PCA is usually conducted to build new variables that can represent the characteristics of the data if a lot of variables, usually more than 3, are provided. However, in this project 5, the flow was in backwards: we would like to conduct PCA first whether it is necessary, and then try to figure out its geographical meaning.

1. **Core code**

# imported libraries

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from mpl\_toolkits.mplot3d import Axes3D

from sklearn.preprocessing import StandardScaler

from sklearn.decomposition import PCA

# 3d projection

fig = plt.figure(figsize = [16, 8])

ax = fig.gca(projection = '3d')

ax.scatter3D(x, y, z, c=z, cmap = 'Greens', marker = ".")

# 2d projections

plt.scatter(x, y, c=z, cmap='Greens')

plt.scatter(y, z, c=x, cmap='Blues')

plt.scatter(x, z, c=y, cmap='Reds')

# After PCA

ind\_val = df\_out.drop(['OSM\_ID'], axis = 1).values

dep\_val = df\_out['ALTITUDE'].values

ind\_val = StandardScaler().fit\_transform(ind\_val)

columns = ['LONGITUDE', 'LATITUDE', 'ALTITUDE']

pca = PCA(n\_components = 2)

pc = pca.fit\_transform(ind\_val)

pcdf = pd.DataFrame(data = pc, columns = ['pc1', 'pc2'])

pc1 = pcdf['pc1']

pc2 = pcdf['pc2']

fig = plt.figure(figsize = (8, 8))

ax = fig.add\_subplot(1, 1, 1)

ax.set\_xlabel('Principal Component 1', fontsize = 15)

ax.set\_ylabel('Principal Component 2', fontsize = 15)

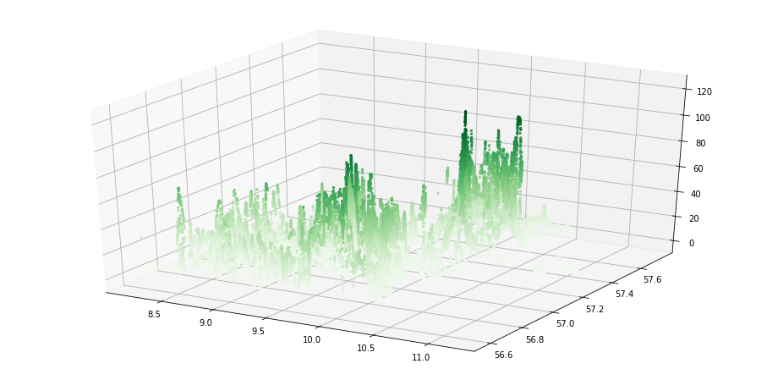
ax.set\_title('2 component PCA', fontsize=20)

ax.scatter(pc1, pc2, c=z, cmap = 'Greens')

ax.legend(['z'])

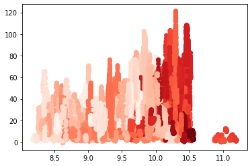
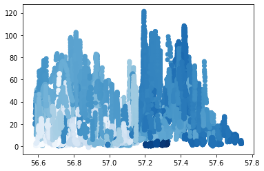
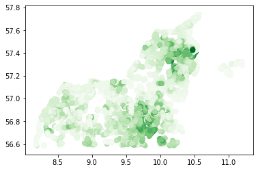
ax.grid()

1. **Results and plots**

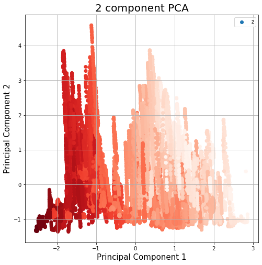
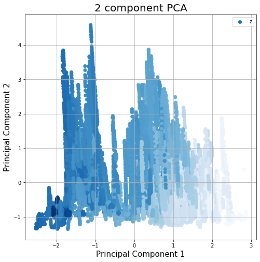
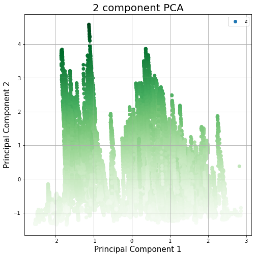


**Figure 1. Plotting of the original data (3D)**

(Data that n(OSM\_ID) < 50 were excluded)



**Figure 2. Plotting of the original data (2D)**



**Figure 3. After PCA (c = based on Altitude, Longitude, and Latitude)**

1. **Discussion**

As we can see in the original plots (Figure 1 and Figure 2), no two components among ‘Longitude’, ‘Latitude’, and ‘Altitude’ cannot tidy up the variation of the other component. For example, the x-y plane projection graph of ‘Longitude’ and ‘Latitude’ cannot show the color gradually, which represents ‘Altitude’.

As we can see in the plot after PCA (Figure 3), in the first plot with green color, two principal components could perfectly explain the gradual transformation of ‘Altitude’. For the other two graphs, we cannot say the two principal components tidy up the variable that represents color. It is more likely to conclude that the PCA would explains the variation of ‘Altitude’ related to its geographic location.

1. **Refernces**

Alexandru-Ionut Petrisor, Ioan Ianos, Daniela Iurea, Maria-Natasa Vaidianu. (2011) *Applications of Principal Component Analysis integrated with GIS*. Procedia Environmental Sciences 14 (2012) 247-256.