Project 3

Optimizing locations of vehicle charging stations, using clustering algorithms.

Roman Kypybida

1. Introduction

   The following project uses the multi-type clustering algorithms to determine the determine the optimal areas for placing vehicle charging stations. It uses a dataset with GPS trajectory data of 10000 taxis in Beijing. Firstly, we use k-means algorithm to divide the area into 15 regions and after perform hierarchical clustering for subdividing the regions and optimizing the travel distance to the charging stations. The weighted k-means is finally performed to find the centers of the extracted clusters, that are to be used as charging stations locations. The results are visualized as maps using "folium" Python package and figures and plots using "matplotlib" package.

2. Theory
   a. K-means algorithms.

      The K-means algorithm chooses random k centers at the beginning of the algorithm and assigns the points to the nearest center. After it computes the real cluster centers and reassigns the points to the nearest new center. It does so until the end of the iteration's procedure or until the exit condition. From the algorithm we can extract the centers of the clusters, clusters labels for the data and "inertia", which is sum of squared distances to the nearest cluster center, weighted, if weights are provided. It can be used to choose the K value using an "elbow" rule.

   b. Hierarchical clustering algorithms.

      The Hierarchical clustering algorithm starts with n clusters, where is n is the number of elements, so at the beginning each element is a cluster center. Then it merges the closest elements into common cluster and does so until one whole cluster is not created.

   c. Weighted K-means algorithms.

      The weighted K-means algorithm weights the elements distances, while computing the clusters, using the provided weights. In the project it is used to compute the charging station location according to the demand.

3. Procedure
   a. Preprocessing data
      i. Dataset

         The dataset contains GPS trajectory data of 10000 taxis in Beijing, in the period of 02 February to 08 February 2008. For each taxi there is a separate file, where each row has id, date, longitude and latitude of the point, where the taxi had been at the specified date.

      ii. Creating a grid

We divide the area between the maximum and minimum longitude and latitude values into a grid, where each cell is 0.005° by 0.005° square.

iii. Choosing the grids

To determine the demand for charging we take out the grid's cells, where the taxis dwelled for more than 20 minutes without entering any other grid cell.

b. Extracting clusters using K-means algorithm.

i. Running the algorithm on the whole dataset

Firstly, we run the algorithm on the dataset, iterating over k value from 1 to n, where n is the number of elements. We extract the value of the "inertia" and save it for further analysis and visualize the results of each step on a map and a figure.

ii. Determining the optimal K value

To determine the optimal K value, we put the values of the inertia on a pot and look for a point, where the inertia had started to become higher, then dropped down and stopped growing. It is supposed to be a point of the highest inertia relative to the other points. Of course, the highest value of the inertia is where the cluster has only one center, but to optimize the number of stations to over one, we look for the value down deeper in the graph.

iii. Extracting the clusters

After we perform the algorithm again and extract the resulting clusters. The resulting clusters centers can be used as locations for the charging stations, but to optimize the number of stations and distance traveled by taxis we continue with the second stage clustering.

c. Multi -type and -stage clustering

i. Performing Hierarchical clustering

1. Running the algorithm

To optimize travel distance of the taxis we perform the hierarchical clustering and subdivide each found cluster into the smaller clusters.

2. Determining the height

Than we look at the graphs and find the optimal cut height for the subdivision. Here the cut height = 0.05.

3. Extracting the clusters

Next we extract the clusters. The resulting clusters in the implementation do not grant us the cluster center, so to determine location of the charging station we move to the further step.

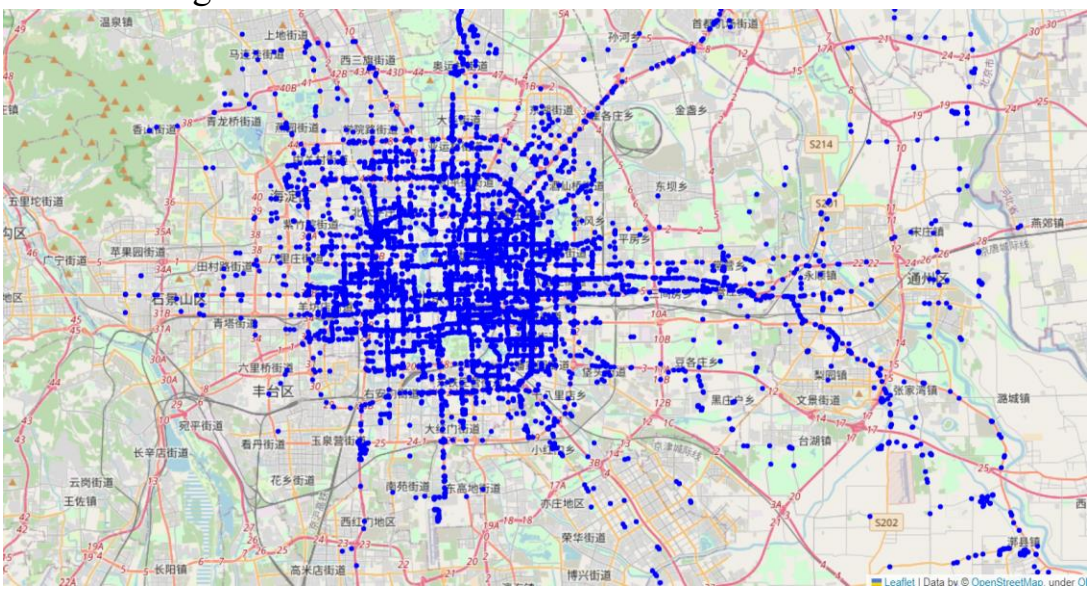
ii. Performing the weighted K-means

To get the locations of the charging stations and account for the demand we perform the weighted K-means algorithm with

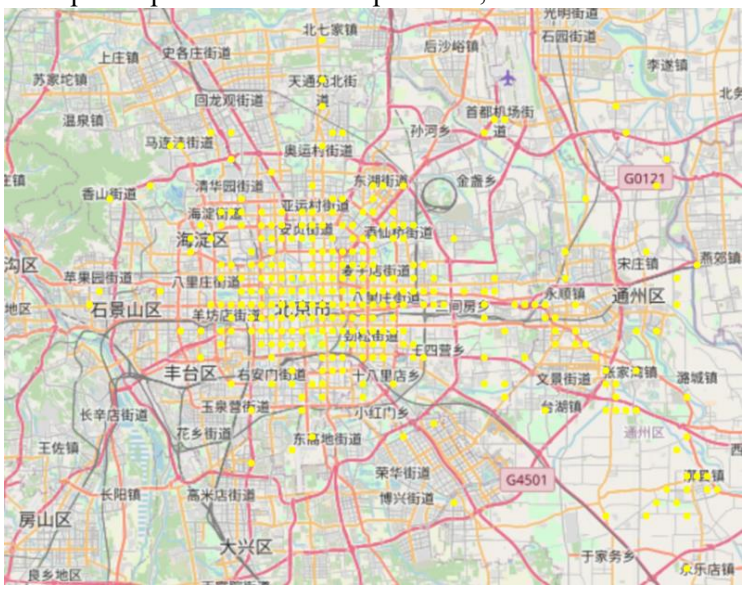
K = 1, weighted by the number of dwelling events (event, when a car dwells for more than 20 min in a cell) and extract the centers and use them as the locations of the charging stations.

4. Results

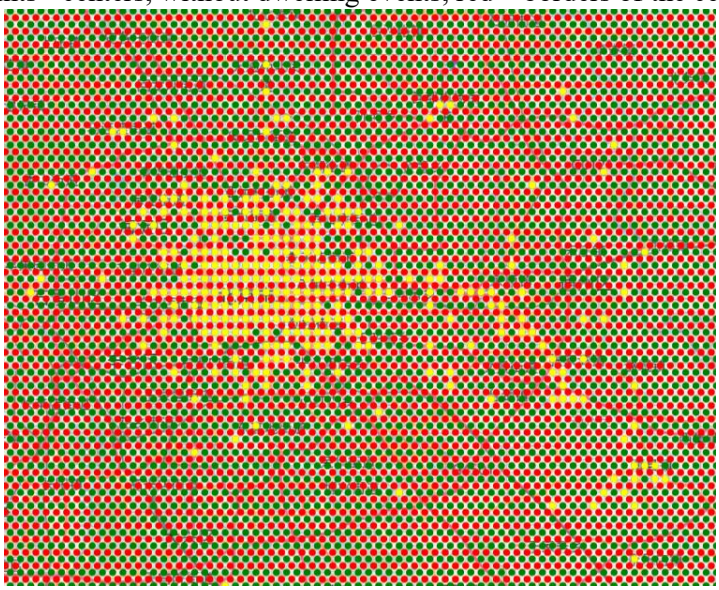
a. The extracted grid



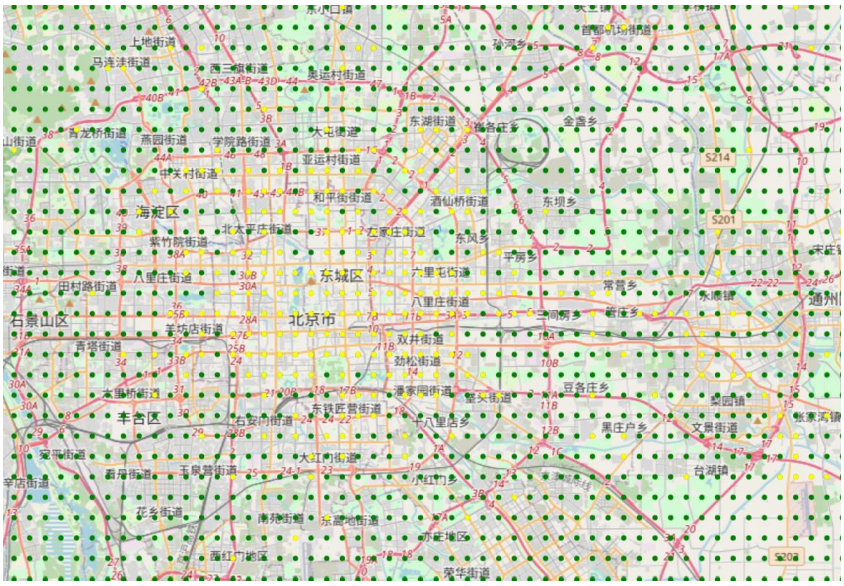
4.1 On this map blue points indicate the positions, where the taxis have been.



4.2 The yellow points show the centers of the cells, where dwelling events occurred. Green points - centers, without dwelling events, red – borders of the cells.

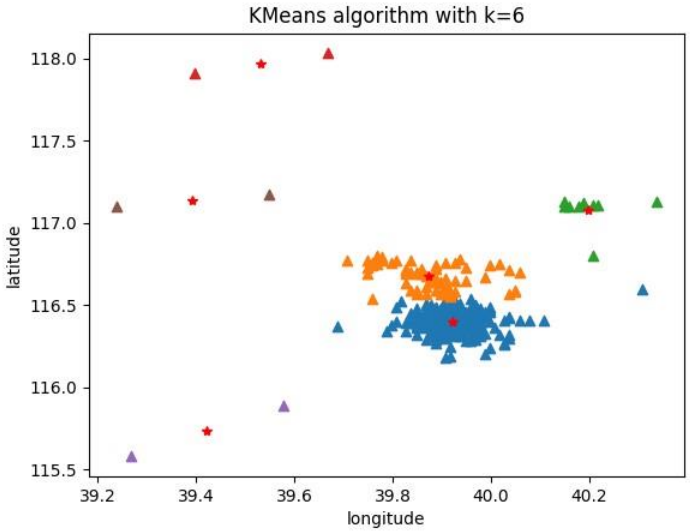


4.3 The yellow points show the centers of the cells, where dwelling events occurred. Green points - centers, without dwelling events, red – borders of the cells.

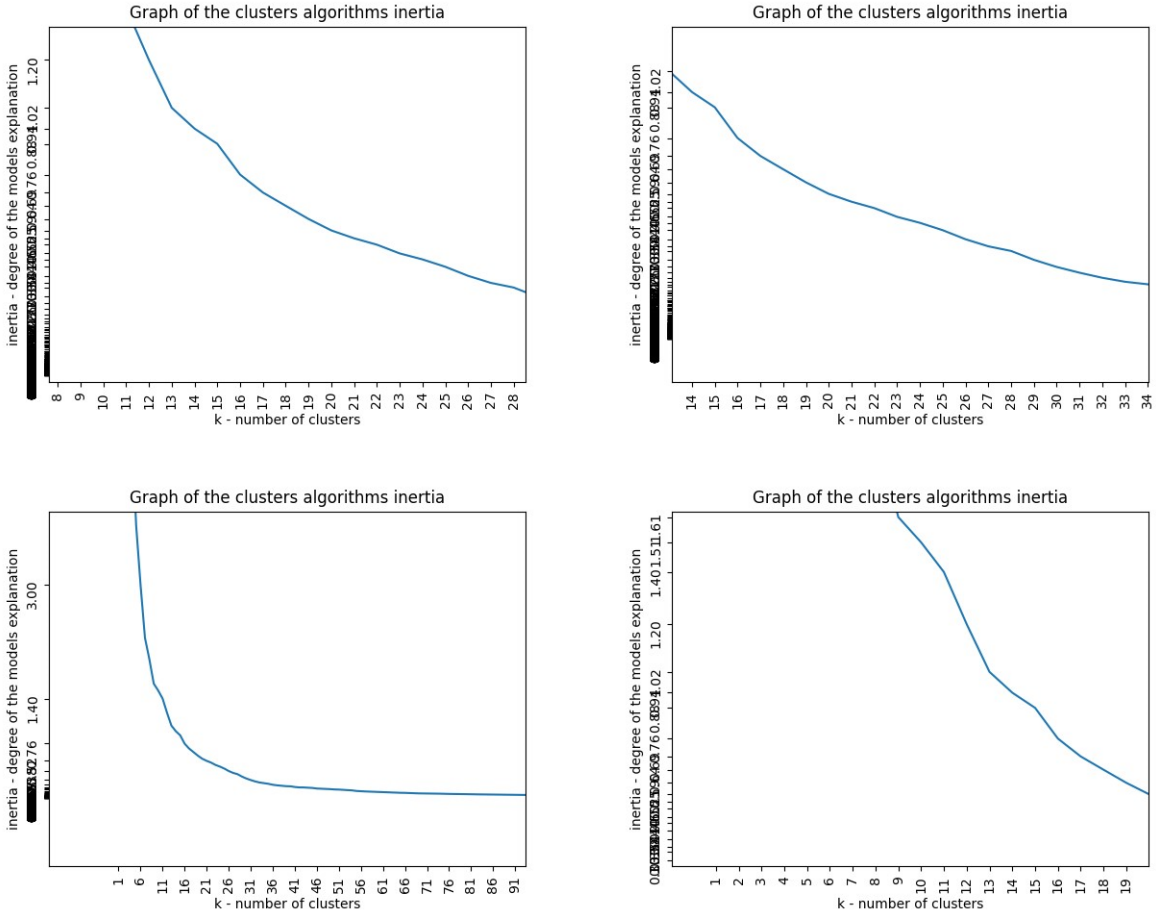


4.4 The same centers together with green points indication other centers, where dwelling events have not occurred.

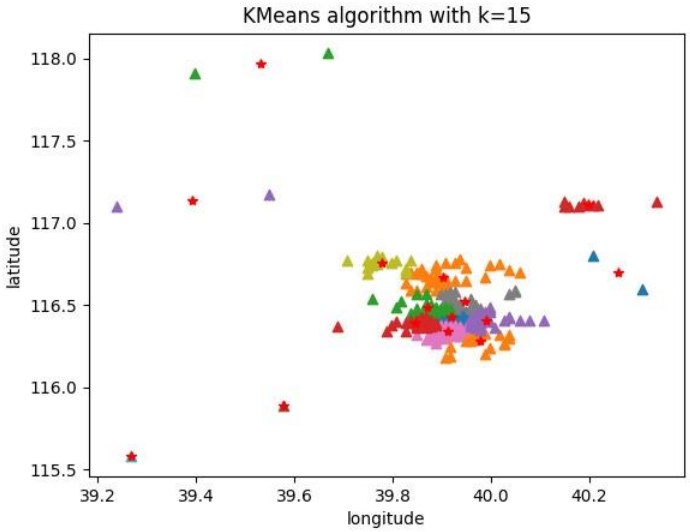
b. K-means algorithm



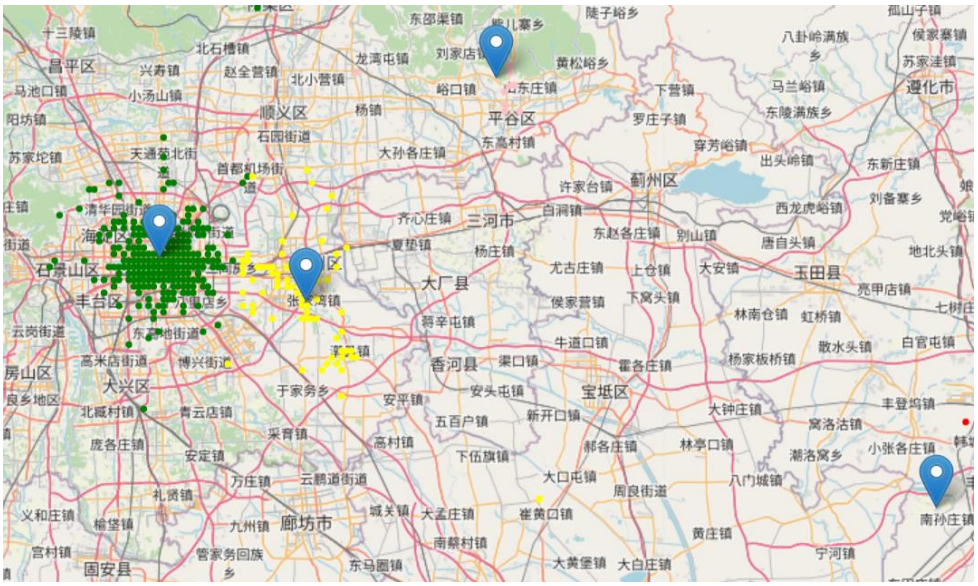
4.5 The K-means clustering with 6 centers.



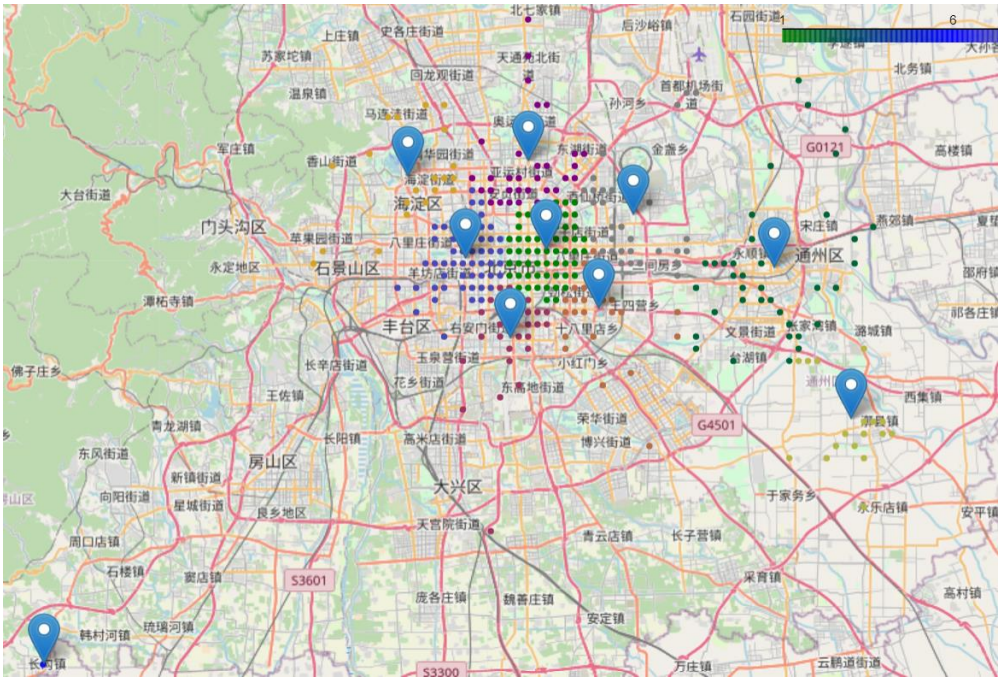
4.6 The plot of the K-means algorithm's inertia against the number of clusters. The K=15 is chosen as an optimal value.



4.7 The K-means algorithm with 15 centers.

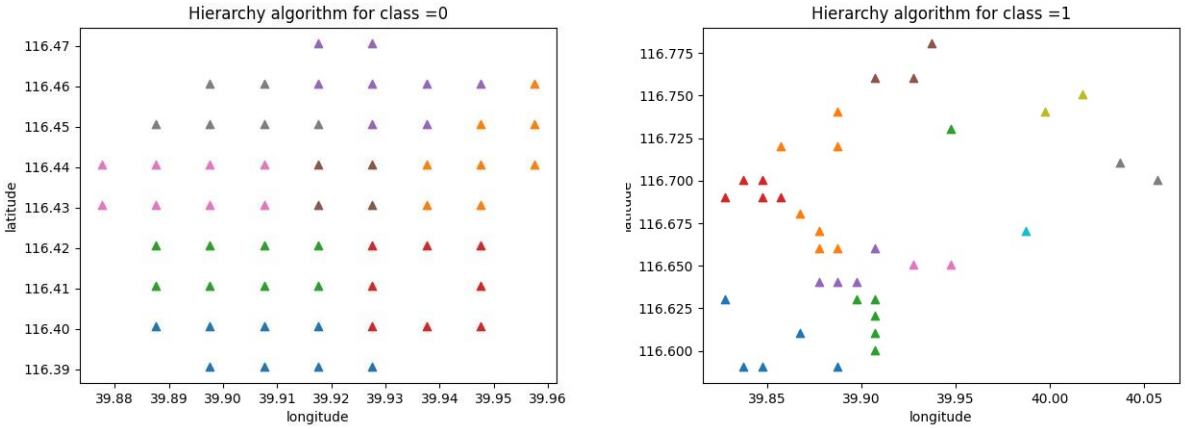


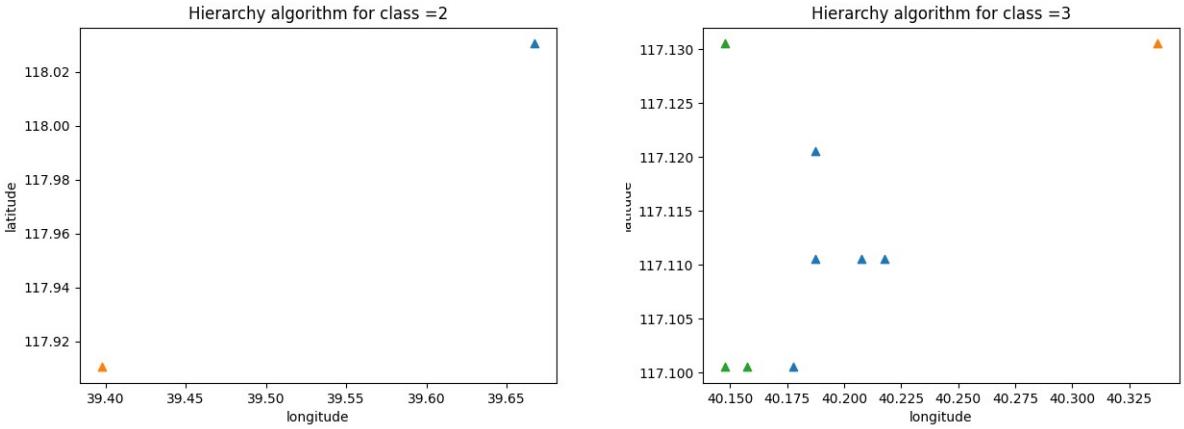
4.8 The K-means algorithms results on a map. K = 4.



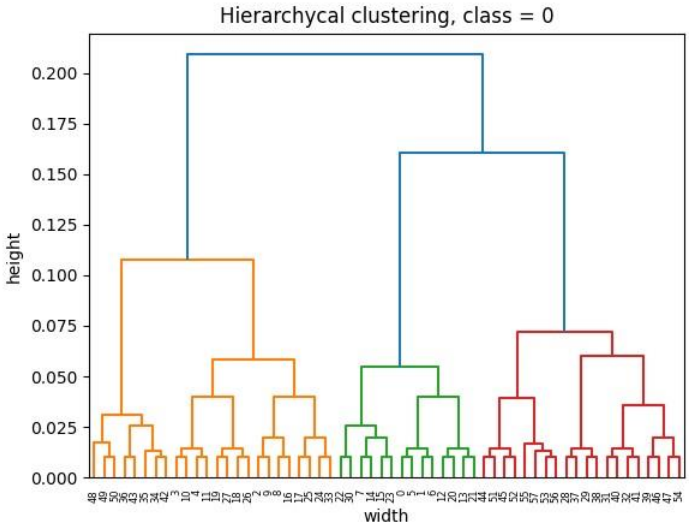
4.9 The K-means algorithm's clustering results on a map. K = 15.

c. Hierarchical clustering algorithm

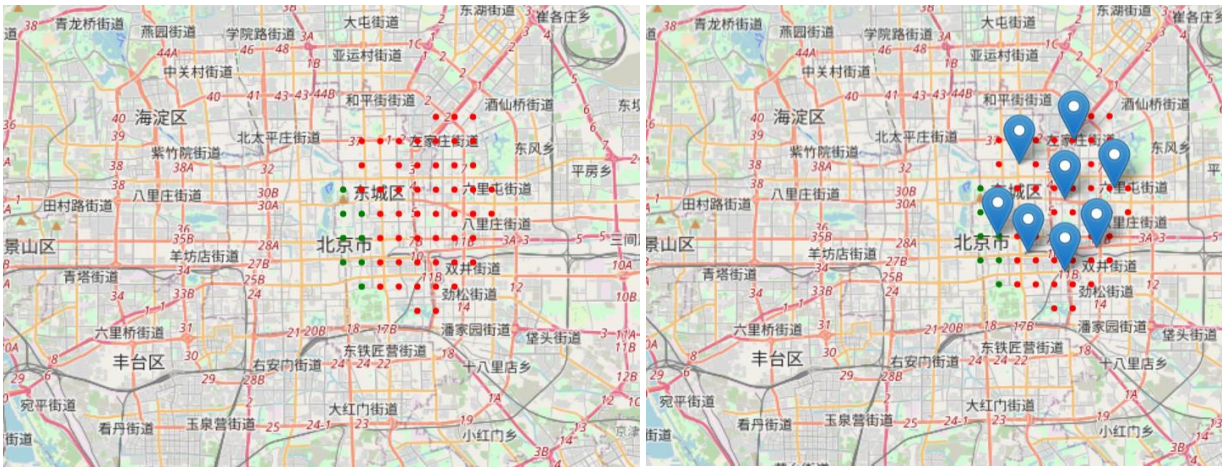




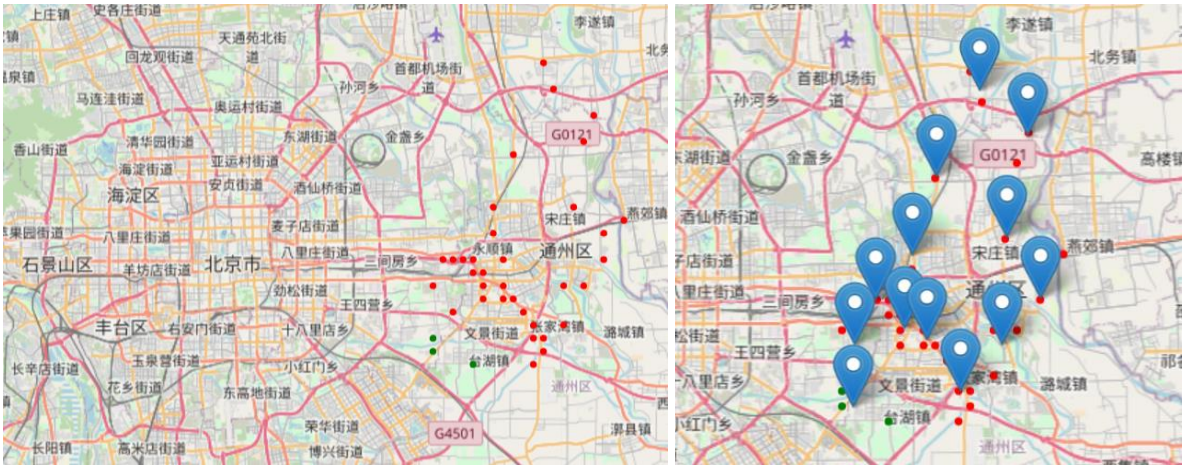
4.10 The plots of clusters for classes 1 – 4 (0 - 3).



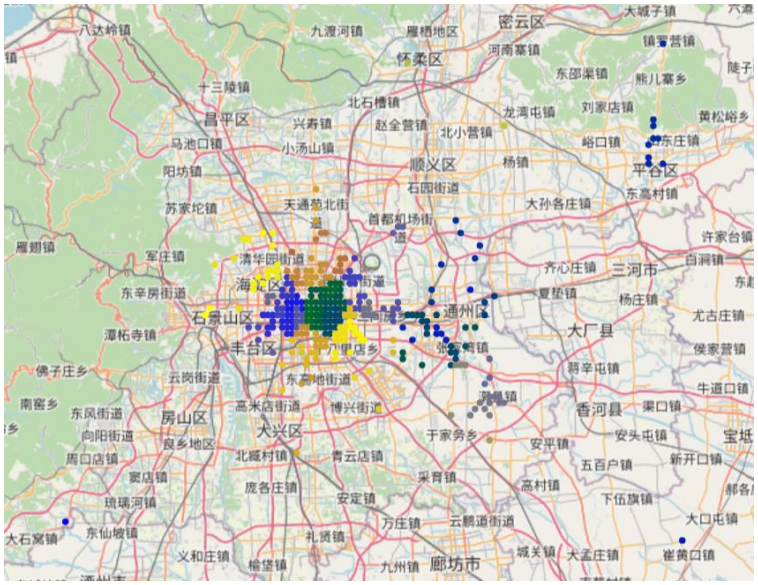
4.11 The dendrogram for class 1 (0).



4.12 The maps of the class 0. Markers indicate the centers of clusters inside the class – locations of the charging stations.

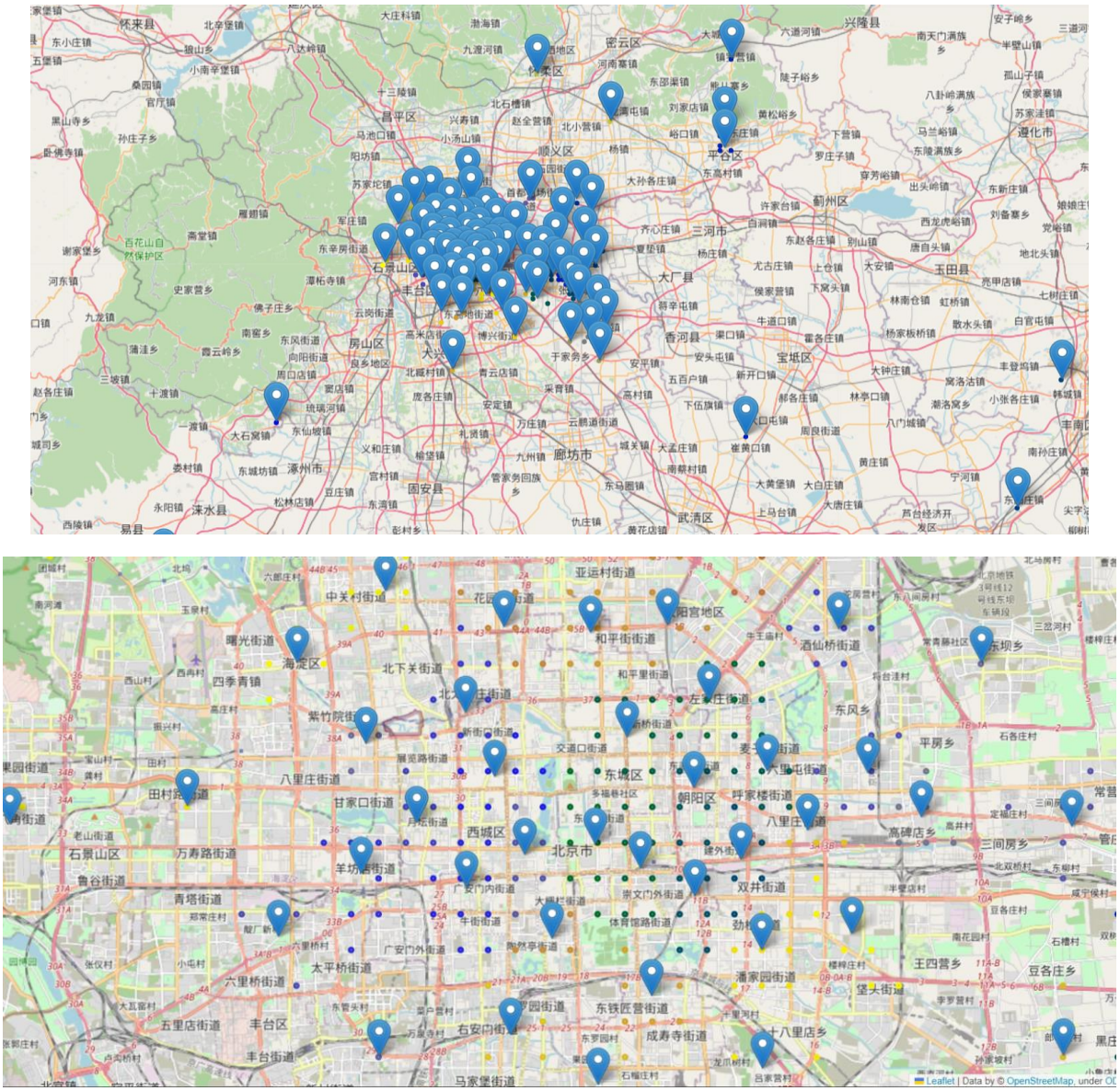


4.13 The maps of the class 1. Markers indicate the centers of clusters inside the class – locations of the charging stations.



4.14 The final division into clusters and subclusters.

d. The weighted K-means



4.15 The final division into clusters and subclusters with markers on the centers of the subclusters.

5. Conclusions

The algorithms have found many point of the suitable locations for the charging stations according to the processed analysis. Although, id does not seem fully optimal in weighting of the places of stations, of which there are too many in the certain locations. This project has used 3-stage algorithm: K-means, Hierarchical, weighted K-means – to optimize number of areas covered by charging stations and locations for the charging stations based on the GPS trajectory of 10000 taxis in Beijing.

References:

[1] Jing Yuan, Yu Zheng, Xing Xie, and Guangzhong Sun. Driving with knowledge from the physical world. In The 17th ACM SIGKDD international conference on Knowledge Discovery and Data mining, KDD '11, New York, NY, USA, 2011. ACM.

[2] Jing Yuan, Yu Zheng, Chengyang Zhang, Wenlei Xie, Xing Xie, Guangzhong Sun, and Yan Huang. Tdrive: driving directions based on taxi trajectories. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS '10, pages 99–108, New York, NY, USA, 2010. ACM