Clustering new real estate properties around metro stations

Georgy Golovanov - 7 June 2020

1. Introduction

Modern cities constantly grow. New properties are being built at any time and require the urban infrastructure to be consistent. This analysis is focused on the real estate properties currently being under construction by clustering them around metro stations to reveal perspective areas. Since the real estate properties development is growing faster than the city infrastructure the analysis can be useful for real estate marketing.

Moscow being a modern permanently growing city is chosen as a subject for the analysis. It has been growing especially fast since 2000s and now combines both real estate development and urban infrastructure improvement. The pattern of the development includes construction activities in suburban areas as well as reconstruction of former factories and manufacture territories within the central city regions.

As a 18 million city, Moscow is confined within approximately 15 km from its center and has a radial structure. Being in a list of the most highly populated cities in the World, it also suffers from the transportation and traffic problems. As a consequence, the fastest and most popular public transport is the underground train system (Moscow metro) with very extensive structure. Since it has a radial structure the metro station density is quite high in city center becoming thin for distant regions.

The main purpose of the analysis is a study of new real estate developments by clustering them around metro stations. The closeness of a metro station becomes crucial for distant and suburban neighborhoods that can be important for real estate marketing estimates. The analysis is based on datasets containing currently developing properties and metro station geolocation.

2. Data acquisition and cleaning

2.1 Data sources

In order to perform the study described above the appropriate datasets have to be obtained, cleaned and modified. The analysis is based on the following data:

1. List of capital construction objects in Moscow;

2. Moscow Metro stations geolocation.

First dataset consists of the full list of ongoing projects currently under development in Moscow. The data are available on the official city administration website as a JSON file (https://data.mos.ru/opendata/7703742961-obekty-kapitalnogo-stroitelstva/passport?versionNumber=1&releaseNumber=886). Among others the fields of interests are property identification number, property region, type, and its geolocation. The full list of objects includes offices and residential facilities.

The geolocation data of metro stations are obtained using Foursquare API. Moscow metro stations are common meeting places, so the Foursquare provides a convenient way of getting their coordinates.

2.2 Data cleaning and manipulation

Both sources has been checked for empty data and no errors were found. The study was focused on residential real estate properties and therefore the construction objects dataset has been filtered in order to keep apartment buildings only. The dataset produced by city government used the Cyrillic encoding (the official paperwork encoding) for the most of the dataset's features. In order to be able to filter by values correctly one has to be applied 'Windows-1251' encoding.

There is a specific Moscow region Zelenograd located far outside the city boundary. This region was removed from the analysis since there was no metro station around. The feature selection for the analysis was mainly caused by the geolocation nature of the study, therefore the coordinates of the object, main function, administrative area it belongs to and id were kept. The coordinate column was split in two (latitude and longitude) columns. Finally, the dataset was represented by 453 objects.

The metro station location dataset was obtained by using '`search`' endpoint of the Foursquare API around the city center. The data in a JSON format were requested and flatten. The shape of the initial data frame was (50 x 18). Despite the fact that Moscow metro system includes about three hundred stations the request output is limited by 50 stations. Only three features were kept for analysis containing station name and its latitude and longitude.

3. Methodology

3.1 Data visualization

The analysis is based on unsupervised machine learning algorithm K-means. The main goal is to form a cluster of developing buildings around a metro station. For that purpose we exploit two methods.

Method 1: the clusters are formed by a single run of K-means algorithm with metro stations geolocation coordinates as fixed initial centroids. It allows to go through a list of metro stations and assign a clusters of buildings to each of them.

Method 2: the clusters are formed by a series of K-means runs with random initial centroids and assign a closest metro station from a list to the found clusters' centroid.

Both methods are used to analyze number of developing properties assigned to each cluster as well as parameters of developing buildings distribution within a cluster such as average distance and variance to the closest metro station.

Figure 1 shows locations of real estate properties while on Fig. 2 the metro stations selected for the analysis are showed.

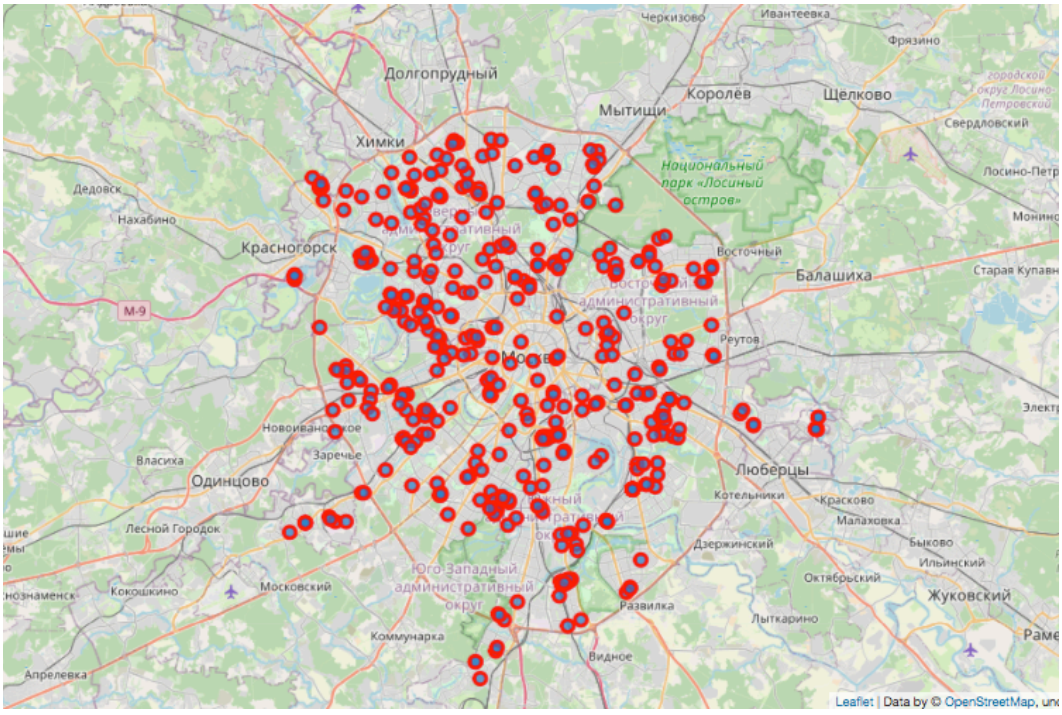

Fig. 1: Location of the developing real estate properties.

The common problem of finding free parameter k is quite straightforward for Method 1, since number of clusters should be equal to the number of metro stations. The second method also can use the same number of clusters for direct comparison between methods, but clustering procedure itself may be more efficient with different number of clusters due to some special features, e.g. two transfer stations from crossing lines can be very close to each others.

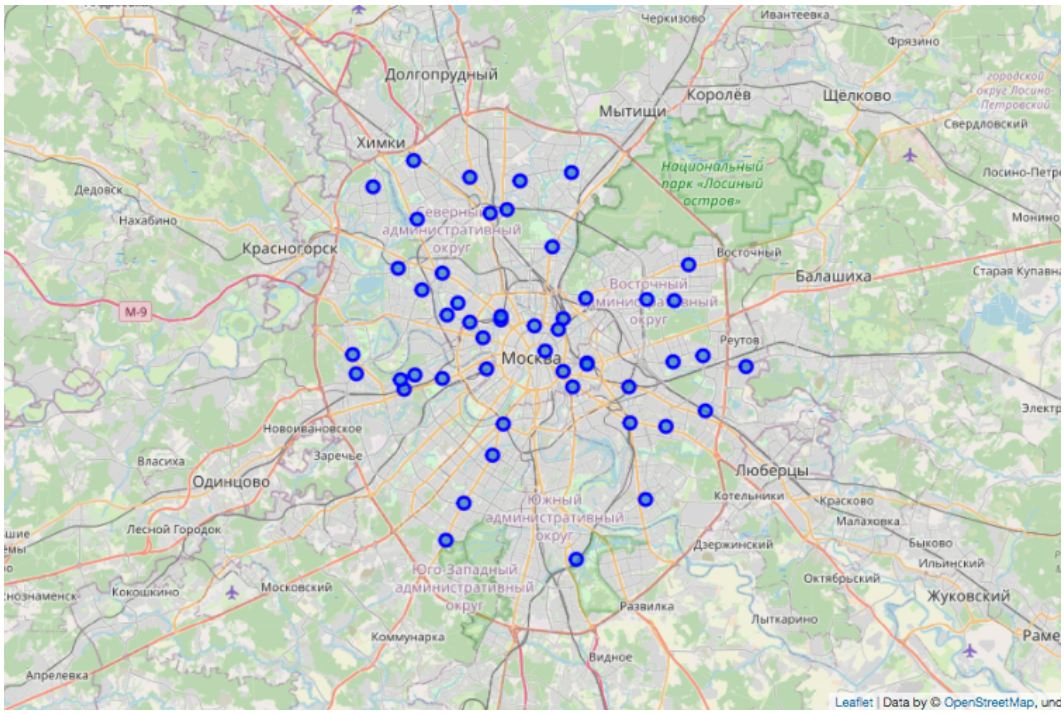

Fig 2: Location of the Moscow metro stations.

3.2 Clustering

The clustering procedure is performed by `KMeans` class from `sklearn` library. As discussed above the centroid initialization is done by assigning metro stations latitude and longitude values. Each cluster is labeled with the same number as the metro station being the centroid of the cluster. Figure 3 illustrates metro stations (blue rounds) and clusters (colored rounds) formed around each of them. Only one run is needed since no centroid recalculation is required. The procedure is converged at iteration 14 and inertia is found to be 0.129.

The second method uses the same approach with the only difference of random centroids initialization, while the number of clusters is kept the same. After formation of new clusters the geolocation coordinates of new centroids are available and therefore one can find the closest metro station using the Foursquare

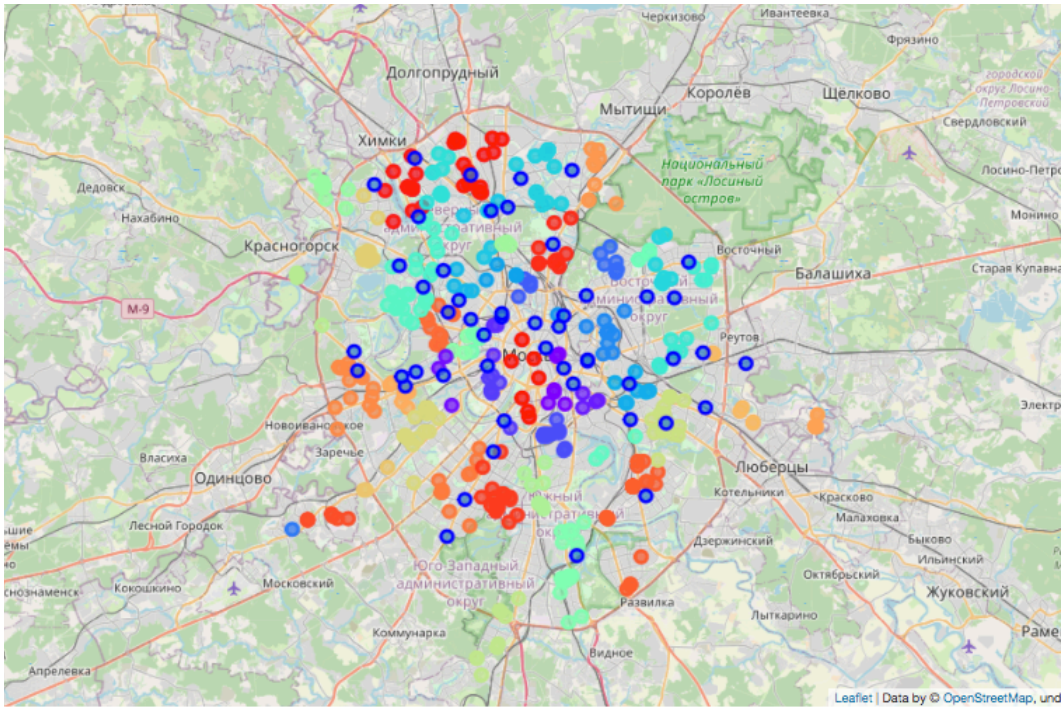

Fig 3: Metro stations and clusters formed around them using the fixed initialization method.

API request. The clusters are shown on Fig. 4 by colored rounds and the found metro stations by blue rounds. The procedure is converged at iteration 11 with inertia 0.103. One can see more sophisticated clustering as compared to the fixed initial centroids.

4. Results

Based on the cluster information one can calculate the number of developing buildings within each cluster as well as average distance to the metro station. This can be very useful for real estate marketing to extract value from the certain properties. A number of buildings for each cluster is shown on Fig. 5 and found to be 9.06 ± 6.72 on average.

A distance between each building within a cluster and corresponding metro station can be calculated for each cluster. Figure 6 shows the boxplot corresponding to a average distance of buildings and variance within each cluster. One can see that the average distance doesn't exceed few kilometers and there are some clusters that have developing buildings within 1 km away from metro stations. From the other hand there are several clusters with average distance more than 10 km. They correspond to the neighborhoods located quite far from the city center that don't have a metro station close to them.

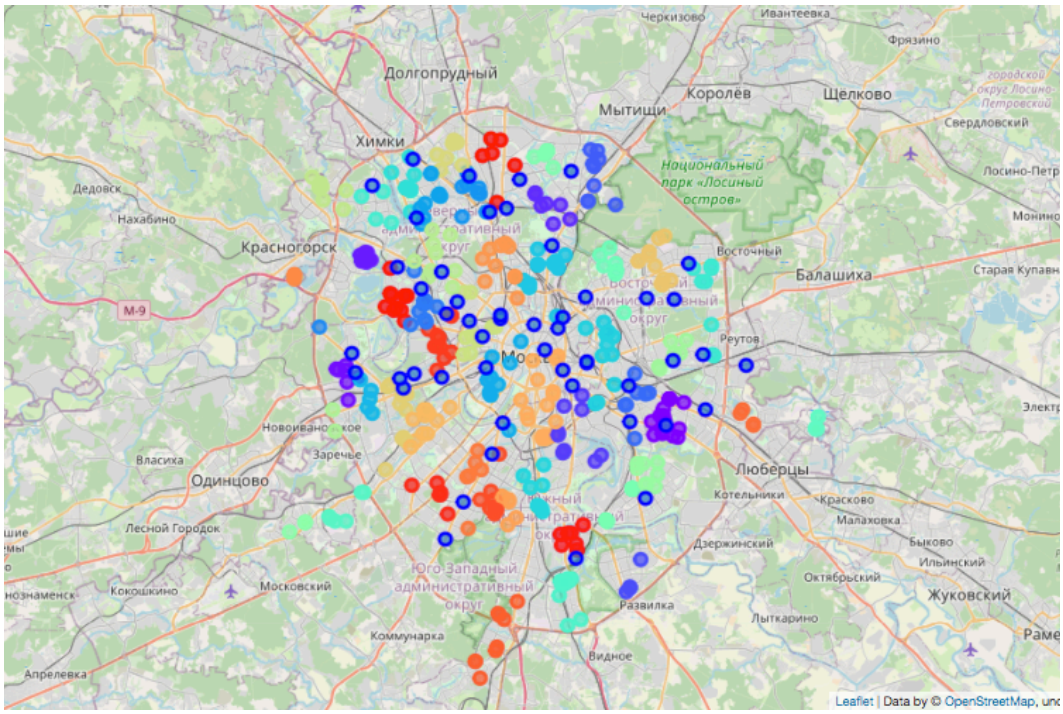

Fig 4: Metro stations and clusters formed using the random initialization method.

5. Discussion

The described procedures perform the first approximation of the new building clustering problem and have some downsides. One of the limitation is that it doesn't take into account closeness of the adjacent metro stations and therefore there is a need of algorithm

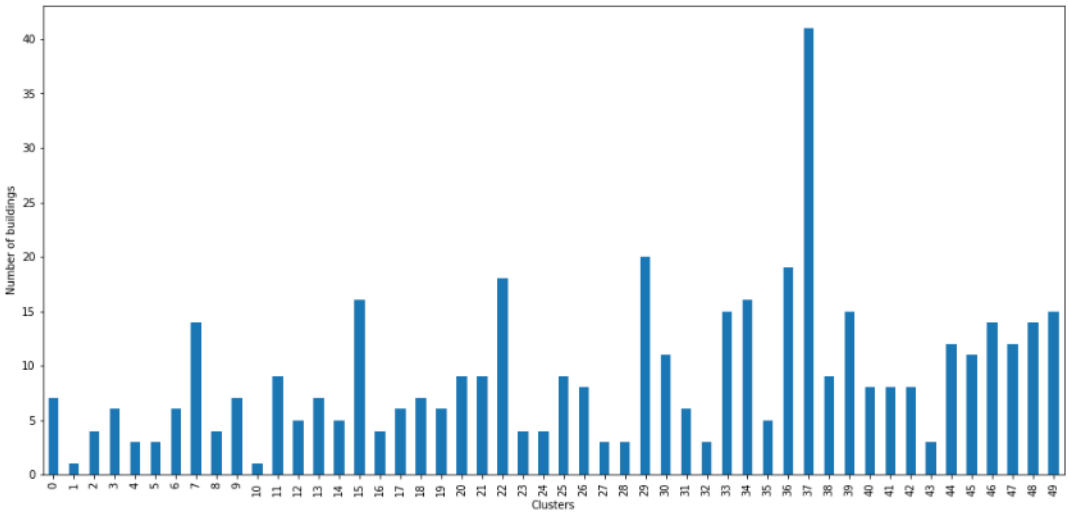

Fig. 5: Number of buildings in each cluster

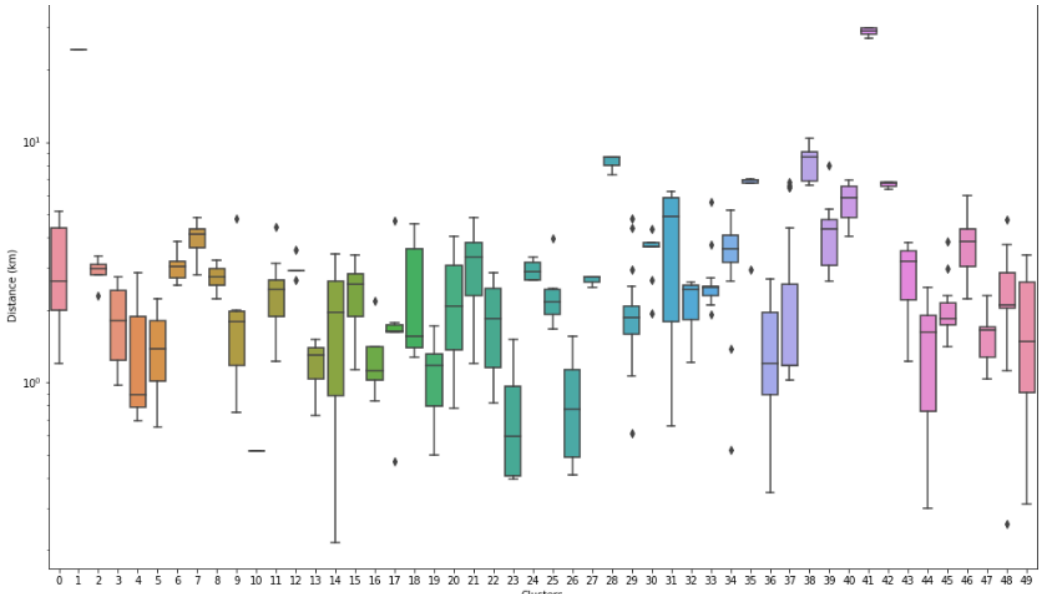

Fig. 6: Distance (in km) between metro station and buildings for each cluster.

improvement in order to assign buildings to a particular station. A fixed number of clusters being an advantage of the analysis, could be a downside of the algorithm at the same time since the adjacent stations have to be a centroids of their own clusters. This leads to a splitting of a good cluster into sub-clusters with uncertain properties. Therefore a clustering of adjacent metro stations into one can help improvement of the algorithm.

The methods used in the analysis also suffer from the limitation of the number of metro stations provided by Foursquare. Using the full list of stations will improve performance since it will correspond to the real coverage which is more dense comparing to the one obtained for the analysis.

In order to calculate distance between two points the Haversine distance was used. This seems to be correct but quite excessively within the current study since there is no need to take into account Earth sphericity.

A possible advantage of the analysis could be a usage of a list of projected metro stations that opens a possibility to estimate values of developing buildings in the near future.

6. Conclusion

The performed analysis is focused on the real estate properties currently being under construction by clustering them around metro stations and shows that one

can reveal perspective areas based on the calculated distance to a nearest metro station. Since the real estate properties development is growing faster than the city infrastructure the analysis can be useful for real estate marketing needs.