SOME ASPECTS OF MACHINE LEARNING IN LOCATION TASKS

Kateryna Kononova[1]

[1]Department of Economic Cybernetics and Applied Economics, V. N. Karazin Kharkiv National University, e-mail: kateryna.kononova@karazin.ua

Abstract. As a result of the study, locations for Pan-Asian food delivery service in Kharkiv have been found so that their network evenly covered the entire city; and different units were at an acceptable distance from each other. The company's order database allowed us to apply ML algorithms, in particular, clustering methods to find optimal locations. Three clustering models were developed and a series of experiments were conducted with each of them. The analysis of the model results allowed us to confirm both hypotheses put forward in the paper, namely: 1) reducing dimension does not skew clustering results obtained on the full database; 2) urban traffic has a significant impact on clustering results. This made us recommend pre-group the data and consider urban traffic in location tasks for the referred company.

Keywords: Location task, Machine Learning, Clustering, Shift Means, K-means, API, Google maps.

CEUR-WS.org/Vol-2649/paper5.pdf

1. Introduction

The company's success significantly depends on the location. It affects not only the cost of rent, access to materials, workers, transportation, but also the perception of the brand and expansion of the customer number.

Location databases have enabled companies to do initial screening themselves, hence reducing their need to rely on external experts to providing only very specific information on locations [7]. Machine Learning (ML) algorithms are effectively used for finding the right locations using accumulated companies' data; especially, clustering methods, which within the geomarketing approach, use spatial data (coordinates, address, registry or other bindings) along with general information.

Various theoretical aspects of ML application in the location tasks are explored in the scientific literature. Montejano et al. overviewed different location models used within the geomarketing field, exemplifying it through the use of Geographic Information Systems (GIS) [8]. Serajnik et al. performed the statistical analysis, evaluated geodata and carried out spatial analysis with a subsequent cartographic visualization to define mall location strategy [11]. Rosu et al. used quantitative models for measuring accessibility to the existing shopping centers in the city, calculating thus their catchment area, for identifying a suitable location for a new shopping center [9].

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

A lot of papers are devoted to the location task analysis in food retail. To propose new locations for the supermarkets, Grassi [6] analyzed four conditions that are prioritizing for new locations: the supermarket area of influence, road access, competition and income of the target audience. Based on the analytic hierarchy process method and spatial analysis of GIS, Xiao [14] built a methodology for the process of selecting a supermarket site location. Bekti [3] and Baviera-Puig provided an application of the spatial methods for retail marketing strategy development for the supermarkets [2].

Geomarketing algorithms are also widely used in international and domestic business practices. For example, WIGeoGIS [13] helps to choose the best GIS system and the relevant market data, as well as implements of mapping solution for location analysis. Ukrainian company GeoDesign [4] offers a business strategy development using spatial data.

2. Hypothesis

The object of the study is the Pan-Asian food delivery service, which is tasked with finding optimal locations for a network of its units. The company is represented in several cities of Ukraine, including Kharkiv where it already has three divisions.

The company is rising rapidly, the number of its customers is growing, so the current production capacity is no longer sufficient to meet the orders flow in strict delivery time limits. Therefore, the company has a need to open two new divisions so that the load at all units was uniform, and the delivery time took no more than 15 minutes. It was decided to renovate the company structure completely closing the old units and opening new ones in the optimal locations.

Delivery time is a key location factor for this company. Thus, it was important to test the impact of urban traffic on model results.

To find the optimal locations, data on the orders made in Kharkiv last year at peak hours (from 15-00 to 21-00) were collected. During this period 36095 orders were received from 9002 customers (table 1).

Table 1. A fragment of the dataset

Order Time	Latitude	Longitude
16:12	50.027284	36.225768
16:27	50.053163	36.197766
15:12	50.019159	36.222417
15:39	50.013091	36.278568
17:53	49.953651	36.214925
15:13	50.013091	36.278568
15:38	49.958041	36.329566
17:21	49.981018	36.147102
17:08	49.950979	36.163038
15:36	49.963422	36.287482

To find dense areas of the customers, we decided to design clustering models.

Since it was decided to use Google Maps API [5] to consider urban traffic in the model, it was crucial to optimize the number of requests sent online at each stage of the clustering algorithm. Hence, it was necessary to choose the method of combining points into a cluster, and decide if it should be based on pairwise distances or centroid method. The advantage of the first-type methods is that they do not need recalculating distances every time after combining, which significantly reduces the computational complexity of the algorithm.

However, according to preliminary estimates, the use of pairwise distances methods requires more than 81 million requests for a base of 9,000 clients; while for the five clusters detecting, centroid methods require about 45,000 requests for each iteration (and given that 30-40 iterations are needed for the algorithm convergence, we get about 1.5 million requests only). Thus, the centroid method has been chosen.

Nevertheless, the question of query optimization remained open. To solve this problem, it was decided to test the hypothesis that dimension reduction does not skew the results of clustering.

Thus, the following two hypotheses were put forward for consideration in the paper:

1) reducing dimension does not skew clustering results obtained on the full database;
2) urban traffic has a significant impact on clustering results.

Three clustering experiments were performed to test these hypotheses:

– full sample clustering,
– pre-grouped sample clustering,
– clustering based on the urban traffic data.

3. Full sample clustering

Mean Shift algorithm based on the centroid method was selected for clustering. Mean shift clustering is a sliding-window-based algorithm that attempts to find dense areas of data points. It is a centroid-based algorithm meaning that the goal is to locate the center points of each class, which works by updating candidates for center points to be the mean of the points within the sliding-window. These candidate windows are then filtered in a post-processing stage to eliminate near-duplicates, forming the final set of center points and their corresponding groups [10].

To use this algorithm with geographical coordinates, one has to select a distance metric. Since the size of the sliding window was given in kilometers, it was decided to use the Haversine metric [12]:

$$d = 2r \cdot sin^{-1}\left(\sqrt{sin^2\left(\frac{x_1 - x_2}{2}\right) + cos(x_1)cos(x_2)sin^2\left(\frac{y_1 - y_2}{2}\right)}\right),$$

where d is the distance between points (in km);
r is the globe radius (6371 km);
x_1, x_2 is the longitude of two points;
y_1, y_2 is the latitude of two points.

The results of the baseline clustering model obtained on the full dataset are presented in Table 2 and Figure 1.

Table 2. Results of the baseline clustering model

Cluster	Center Coordinates	Number of customers	Customer share	Number of orders	Share of orders
1 – blue	49.9531, 36.2977	1589	17.65%	5426	15.03%
2 – red	49.9879, 36.2218	2766	30.72%	10148	28.11%
3 – green	50.0430, 36.2229	1633	18.14%	7917	21.93%
4 – pink	49.9377, 36.3872	765	8.49%	2326	6.44%
5 – orange	50.0117, 36.3407	2249	24.98%	10278	28.47%

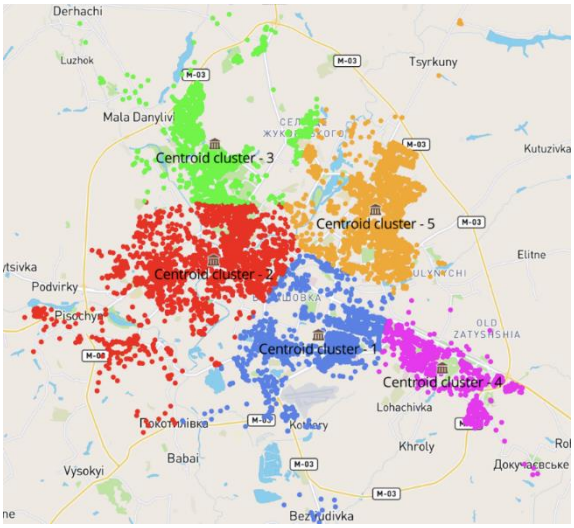

Fig. 1. Visualization of the baseline clustering model

4. Pre-grouped sample clustering

To check the first hypothesis, a weighted clustering method was used to reduce the dimension. A centroid calculated by a weighted value considers each customer to have individual value. The centroid is not created in the center of all customers but in the center of the customers who most satisfy the value, one has weighted [1].

Using a k-means algorithm with the Haversine metric to detect 500 clusters, the coordinates of the weighted centers were obtained.

Figure 2 presents the initial points (marked red), and the weighted centers (marked purple).

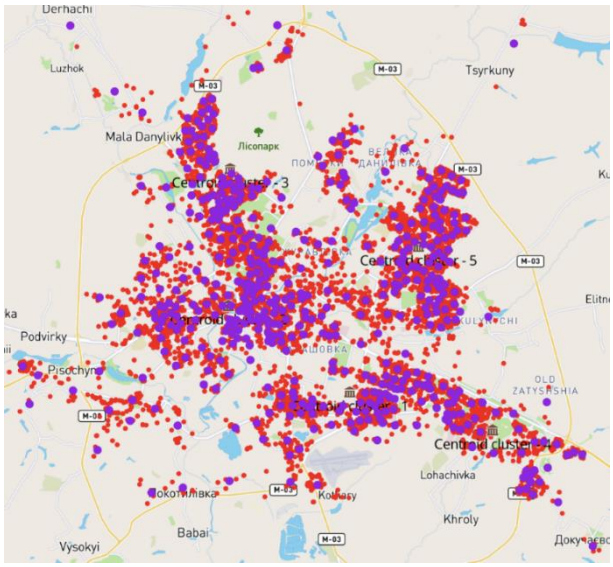

Fig. 2. Visualization of clients' dataset after dimension reduction

Next, the set of 500 points was clustered using the Mean Shift algorithm and the following estimates were obtained (Figure 3, Table 3).

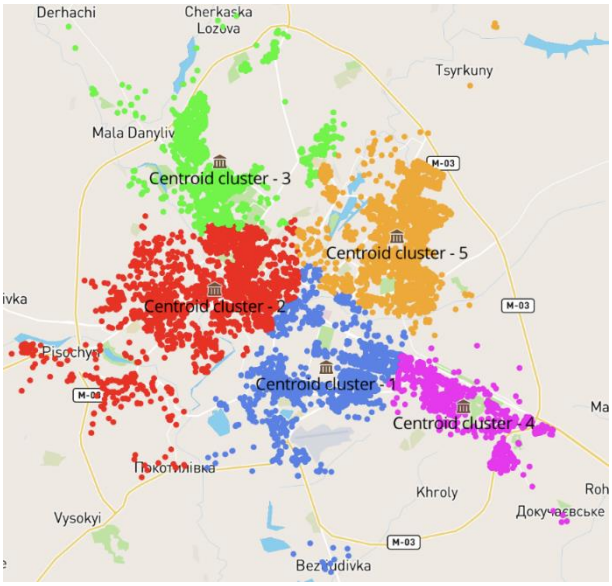

Fig. 3. Visualization of the pre-grouped sample clustering

Table 3. The results of the pre-grouped sample clustering

Cluster	Center Coordinates	Number of customers	Customer share	Number of orders	Share of orders
1 – blue	49.9554, 36.2939	1633	18.14%	5802	16.07%
2 – red	49.9887, 36.2196	2692	29.90%	9744	26.99%
3 – green	50.0434, 36.223	1612	17.90%	7818	21.65%
4 – pink	49.9386, 36.3851	817	9.07%	2444	6.77%
5 – orange	50.0115, 36.3408	2248	24.97%	10287	28.49%

Comparative analysis of the first and second clustering results showed that the reduction of dimension does not lead to its significant skew, the obtained clusters coincide by 97%. This allowed us to use the pre-grouped dataset for the calculations, which require urban traffic data.

5. Clustering based on the urban traffic data

To consider urban traffic data, the clustering algorithm has been modified – to measure the distance between the two points instead of the Haversine metric we used data provided by Google Maps [5].

Working with the Google Maps API, the following settings were specified: type of transport was ‘car’; forecast time was ‘02.12.2019 18:00’; forecast type was ‘most likely’ (‘pessimistic’ and ‘optimistic’ estimates were also tested).

As a result of the clustering model with regard to urban traffic, the following estimates were obtained (Figure 4, Table 4).

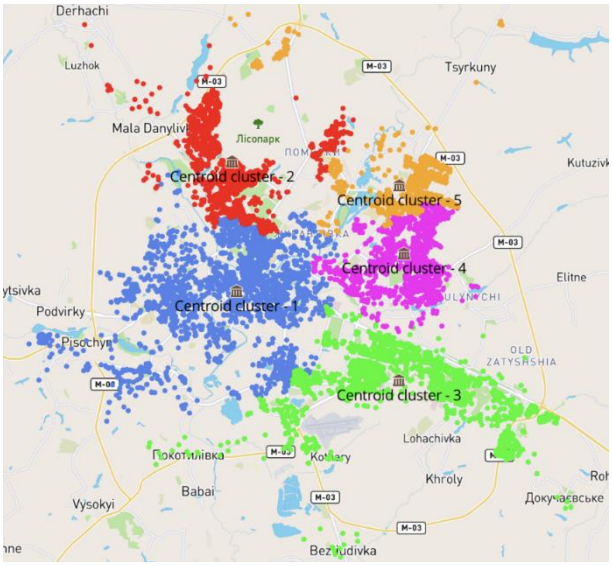

Fig. 4. Visualization of clustering with regard to urban traffic

Table 4. Clustering results with regard to urban traffic

Cluster	Center Coordinates	Number of customers	Customer share	Number of orders	Share of orders
1 – blue	49.9844, 36.2261	3109	34.53%	11468	31.77%
2 – red	50.0411, 36.2229	1608	17.86%	7763	21.50%
3 – green	49.9452, 36.3366	1943	21.58%	6328	17.53%
4 – pink	50.0009, 36.3401	1465	16.27%	6578	18.22%
5 – orange	50.0307, 36.3369	877	9.74%	3958	10.96%

Comparative analysis of three clustering experiments showed that traffic data significantly affects the clustering results. This leads to the conclusion that it is necessary to consider this factor in location tasks for the referred company.

6. Conclusions

As a result of the study, locations for Pan-Asian food delivery service in Kharkiv have been found so that their network evenly covered the entire city; and different units were at the acceptable distance from each other (less than 15 minutes by car).

To find the locations, data about 36095 orders from 9002 customers were collected. The location database has enabled us to do screening using ML algorithms, in particular, clustering methods, which within a geomarketing approach, use spatial data along with general information.

Two hypotheses have been put forward for consideration, namely:

1) reducing dimension does not skew clustering results obtained on the full database;

2) urban traffic has a significant impact on clustering results.

Three clustering models were developed and a series of experiments were conducted with each of them.

Comparative analysis of the first and second clustering results showed that the reduction of dimension does not lead to its significant skew. This allows us to use the pre-grouped dataset for the calculations based on urban traffic data.

To calculate the distance between two points with regard to urban traffic, the Haversine metric has been replaced with Google Maps data. The analysis of the experiments showed that traffic data significantly affects the clustering results.

Thus, as a result of the study, both hypotheses were confirmed. This made us recommend pre-group the data and consider urban traffic in location tasks for the referred company.

References

1. ArcGIS. (2019). Find optimal store locations. Retrieved from http://desktop.arcgis.com/ru/arcmap/latest/extensions/business-analyst/find-optimal-store-locations-mean-store.htm.

2. Baviera-Puig, A., Buitrago-Vera, J., & Escriba-Pere, C. (2016). Geomarketing Models in Supermarket Location Strategie. Journal of Business Economics and Management, 17(6), 1205–1221. DOI: 10.3846/16111699.2015.1113198.
3. Bekti, R., Pratiwi, N., & Jatipaningrum, M. (2018). Multiplicative Competition Interaction Model to obtained Retail Consumer Choice based on Spatial Analysis. IOP Conference Series: Earth and Environmental Science, 187(1), 1-9. DOI: 10.1088/1755-1315/187/1/012041.
4. Geodesign.info. (2019). Retrieved from https://geodesign.info/.
5. Google Maps. (2019). Retrieved from https://www.google.com/maps/.
6. Grassi, V. (2010). Estratégias de localização de uma rede de supermercados: o geomarketing aplicado à companhia zaffari em Porto Alegre. Porto Alegre, Brasilia: Universidade Federal do Rio Grande do Sul. DOI: 10.13140/RG.2.2.23534.31041.
7. Heil, K. (2012). Location strategy. Retrieved from https://www.referenceforbusiness.com/management/Int-Loc/Location-Strategy.html.
8. Montejano, J.A., & Cruz Bello, G.M. (2018, February 5). Geomarketing Localization Models. Espacialidades, Revista de temas Contemporáneos sobre lugares, política y cultura, 8(1), 95-120. Retrieved from http://espacialidades.cua.uam.mx/vol/08/2018/01/06_Montejano_y_Cruz.pdf.
9. Rosu, L., Blăgeanu, A., & Ionuţ-Ciprian, I. (2013). Geomarketing. A New Approach in Decision Marketing: Case Study – Shopping Centres in IASI. Lucrările seminarului geografic 'Dimitrie Cantemir', 36, 123-133. Retrieved from https://www.researchgate.net/publication/291956981_ GEOMARKETING_-A_NEW_APPROACH_IN_DECISION_ MARKETING_CASE_STUDY_-_SHOPPING_CENTRES_IN_IASI.
10. Seif, G. (2018). The 5 Clustering Algorithms Data Scientists Need to Know. Retrieved from https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68.
11. Serajnik, T., Amaduzzi, S., & Paulus, G. (2014). Geomarketing. Analyses of the Città Fiera Ma. GI_Forum, 1, 105-114. DOI:10.1553/giscience2014s105.
12. Sinnott R.W. (1984). Virtues of the Haversine. Sky and Telescope 68 (2), 159.
13. Wigeogis. (2019). Transparency and success. Geomarketing supports retailers. Retrieved from https://www.wigeogis.com/en/retail_geomarketing/.
14. Xiao, D., & Ye, W. (2019). Combining GIS and the Analytic Hierarchy Process to Analyze Location of hypermarke. IOP Conference Series: Earth and Environmental Science, 237, 1-5. DOI: 10.1088/1755-1315/237/3/032012.