Service Point Placement By Customer Geo-Location Clustering

Location, Location, Location When Distance, Distance, Distance Matters

Constantia Malekkou
Data Science Project
University of Nicosia
Nicosia, Cyprus
constantia.malekkou@gmail.com

ABSTRACT

This project has been developed for COMP592DL - Project in Data Science, for the MSc in Data Science at the University of Nicosia.

The purpose of this project is to try to identify the optimal number and placement of physical store locations, for a retailer company that depends on a physical network for the distribution of goods and services to its customers. The target variable is the minimization of the distance to the nearest service point by the retailer's customers.

For this project, we have gathered open-source data to serve as a sample population, and used the k-Means unsupervised clustering algorithm, for the model development. The results of the model are the optimal number and geographical positions of service points for a retailer company, based on the geolocations of the retailer's customers. The service points proposed by the developed model minimize the distance-to-be-travelled by the customers to the nearest service point, with the least number of service points, resulting to reduced operating costs.

KEYWORDS

Geolocation, Location, Address, Service Points, k-Means Clustering, Big Data

INTRODUCTION

The location of a physical store is important for any retailer company that relies on a distributed branch network for the promotion of its goods and services. The distance to the nearest service point is an important factor to consider, as it is a measure of convenience for its customers. The initial set-up costs required for the establishment of a new service point should be carefully weighed against the expected benefits [1], as future relocations or new establishments translate to additional, unnecessary costs. Such data driven insights would be useful to the retailer company's management to support management decisions regarding the retailer's physical service point locations.

For this project we will be dealing with the case of the service point placement of a banking institution, which was the case examined in [1]. A basic assumption was that the distance a customer was willing to travel was influenced by the population density at the demand point. Customers in high density areas were willing to travel less distance to be serviced, whereas customers in low density areas were willing to travel greater distance to be serviced. Additional literature suggests that bank customers tend to travel to the nearest appropriate banking facility or ATM to be serviced [2]. Thus, locational convenience has always been an objective for banking institutions.

The service point location problem is known as the **Maximal Covering Location Problem** and has been formally defined since 1974 [3]. The measures proposed that would indicate optimal location solutions were: the total weighted distance or time for travel to the facilities, and maximal service distance, which is the distance or time that the most distant user would have to travel to reach a facility. The solution to this problem was to identify service points that would maximize population coverage within a desired service distance by locating a fixed number of facilities.

Various improvements to this problem statement exist in the literature. In [4], the concept of Weighted Benefit Maximal Covering was introduced by supporting the assumption of "step" distances. The number of distance steps could be varied per case examined, and a different weight could be applied to each step to represent the "benefit" in distance for each distance interval. The notion of partial coverage for the Maximal Covering Location Problem was introduced in [5]. The proposed model would classify as fully covered when the service point is found within a minimum critical distance, as partially covered when the service point is found between a minimum and a maximum critical distance, and as not serviced when the service point is beyond the maximum critical distance. In [6], the notion of the service quality was introduced, by replacing the actual distance measurement in the proposed model, with a service level function of the distance to the nearest service point.

The problem with all above solutions and variations to the maximal covering location problem was with the demand representation [7]. The demand, which in our case is the retailer's customers, was represented by a set of discrete demand points with weights on a grid. The differentiation in [7] was to assume uniform distribution of demand points within a specified grid to calculate the population coverage. The results showed significant differentiation in the proposed optimal service point locations.

For the purposes of this project, we have managed to represent each individual demand point without aggregation. This was possible with the translation of sample text addresses to their digital equivalent through utilization of the Google Maps Geocoding API available in Google Cloud Platform. The geographical coordinates were also used for the distance calculation between individual demand points and service points.

The outcome of this project is the identification of the optimal service point number and location with the use of the k-means clustering algorithm on the granular spatial representation of demand. The model's suggested locations have been compared against two sample companies' existing service points. The results of the model show a significant improvement in minimizing the distance to be travelled to the nearest service point, while at the same time proposing a reduced number of service points.

DATA SOURCES

The data source for the model's sample population was the Cyprus National OpenData Portal [8]. More specifically, we have used the following files from the "Μητρώο Εγγεγραμμένων Εταιρειών, Εμπορικών Επωνυμιών και Συνεταιρισμών στην Κύπρο" [9]: "Κατάλογος Οργανισμών", which contains the list of all legal entities of Cyprus that are registered with the Registrar of Companies, and "Κατάλογος Διευθύνσεων Εγγεγραμμένου Γραφείου", which is a list of all the registered office addresses of legal entities ever registered in Cyprus. The two files can be joined together to form one unified dataset based on the address sequence number, present as a key in both files.

The geolocations for the registered office addresses have been found by calling the Geocoding API in Google Cloud Console, with each text address as parameter.

We have gathered information regarding a company's existing service points from a sample company's website. Additionally, the service points from a second competing company, offering the same range of products and services to the same customers, have also been found through the competitor's website. The two sources will be used to compare the existing service point placement of a company versus its competitor. To evaluate the model's performance, the two competing companies' service point placements will be compared against the model service point placement suggestions. The companies selected for this project are banks since they rely on a distributed network of physical branches for customer service, and their service point locations are publicly available. For the rest of this project, they will simply be referred to as CompA and CompB without disclosing any further details.

PREPROCESSING STEPS

The preprocessing steps that have been performed are described below per project area:

1 Addresses

- For the address translation we have dropped the Address Line 2, which contained mostly information on building and/or flat number which are not useful for the address translation
- Addresses that returned 'Null' as geolocations from the Geocoding API were dropped.
- Matching of the unique text addresses with the geolocations from Geocoding API were matched in excel. Further matching of the text address on the addresses starting file to have the information of the addresses seq no, necessary for matching with the customers' file, was also performed in excel.

The population of addresses through preprocessing steps has changed as follows:

Total Initial Addresses	178.916
Unique Initial Addresses Population	124.297
Total Geocoded Addresses	115.191
Unique Geocoded Addresses	79.894

2 Customers

- Filtered for active customers, recognized by their ORGANISATION_STATUS = 'Εγγεγραμμένη'
- Dropped unnecessary customer details
- Dropped customers without a key address sequence number
- Dropped customers with addresses outside the range of Cyprus geocoordinates. Minor note that the valid range used is a best approximation, since the interesting region cannot be exactly defined by rectangular geometry.
- Following investigation, Cyprus lies at a latitude of 34˚33˚ - 35˚34˚ North and longitude 32˚16˚ - 34˚37˚ East approximately. Further adjustment of these coordinates was necessary, to filter out entities registered in the Turkish occupied part of Cyprus unfortunately. This was necessary because distance to the service point is of high importance, therefore addresses outside of a specific range were removed as outliers.

The population of customers through preprocessing steps has changed as follows:

Total Initial Customers	509.647
Total Filtered Customers	238.725
Filtered Customers with Geocoded Addresses	218.738
Filtered Customers with Geocoded Addresses within Cyprus Geocoordinates	217.788

MODEL DEVELOPMENT

For the model development, initially we measured the Haversine distance [10] of all customers to every existing service point of

CompA and CompB. The Haversine distance is formulated to calculate the distance between points on a sphere, given their geographic coordinates of latitude and longitude. Given that the Earth is spherical, the Haversive distance gives a more accurate distance measurement between coordinates, than the Euclidean distance, which measures distances on a flat surface [11].

The service points of CompA were 60, and the service points of CompB were 42. Initially, we calculated the minimum distance of each customer to the nearest service point of CompA and CompB. We then grouped by the nearest derived service point, to calculate the mean, median, min, max, 5^{th} percentile and 95^{th} percentile for comparison purposes with the model's recommendations per service point.

To capture the volume of customers per service point, we also counted the number of customers by grouping by the nearest derived service point for CompA and CompB. Additional metrics for comparison purposes regarding mean and median number of customers were calculated.

For the model development we have used the k-Means clustering method [12]. To determine the optimal number of clusters [13], we have measured the sum of squared distances (error rate) between each data point to its respective cluster center (centroid). The goal of k-means clustering is to minimize the error rate, to identify compact and well-defined clusters. The error rate has been plotted against the number of clusters, in a plot widely known as the "elbow" plot [14].

In the "elbow" plot, as the number of clusters increases, the error rate generally decreases because more clusters provide a better fit for the data points. However, the rate of improvement in the fit will decrease as more clusters are added, eventually reaching a point where adding more clusters does not lead to a significant reduction in the error rate. This is because the additional clusters only capture the noise or minor variations in the data, rather than revealing meaningful patterns. The elbow point is a good trade-off between having a small number of clusters and minimizing the within-cluster sum of squared distances.

A business explanation of the "elbow" plot in this case, is that the number of clusters represents the number of service points to be maintained by the business. The cost-effectiveness of setting up an additional service point should be examined against the benefit gained in terms of the total distance to be travelled to the nearest company's service points [3]. The cost of each service point is made up of the initial setup cost, for example land acquisition, construction, or renovation, and the cost for its ongoing operation, for example rent (if not owned), employee payroll, utilities, and facility maintenance [2].

The error rate in this business setting is the measurement of the total distance to the nearest service point. A reduction of the error rate translates to a reduction in the total distance to the nearest service point of all customers in the dataset. The "elbow" point marks the point that the reduction in the error rate, i.e. total distance to all service points, starts diminishing. Therefore, the cost-effectiveness of adding service points past the elbow point is also diminished.

For the sample model population, there are 2 noticeable "elbow" points – at 31 and 37 clusters, as shown in Figure 1. For our model we have decided to proceed with the more conservative approach and set the optimal number of clusters to 37.

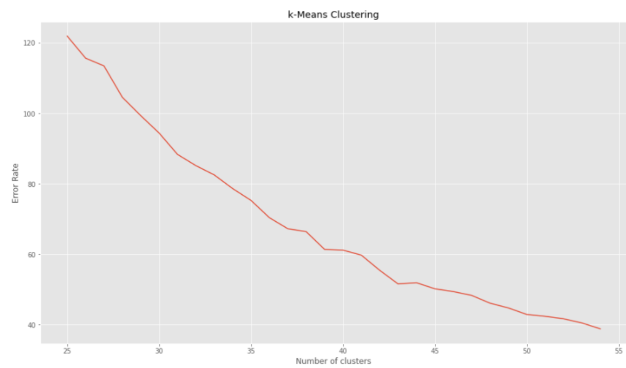

Figure 1 The "elbow" plot of the sum of squared distances (error rate) against the number of clusters for the developed k-Means clustering model

RESULTS

The results of the k-Means clustering model, with regards to the number of service points, showed that the optimal number of service points is 37, compared with 60 and 42 service points of CompA and CompB respectively. Therefore, both companies would initially benefit from decreased operational costs from the reduced number of service points they would have to maintain. The results of the k-Means clustering, with 37 clusters on our dataset, are shown on a map of Cyprus in Figure 3.

For each of the service points we have collected their geographic coordinates, which have in turn been translated to actual Cyprus addresses, as shown in Figure 2.

Service Point	NoOfCustomers	SP_coor	SP Address
0	40647	(35.165411, 33.357317)	5985+53P, Themistokli Dervi, Nicosia, Cyprus
1	9644	(34.695984, 33.031706)	Spyrou Kyprianou Ave 104, Limassol, Cyprus
2	6194	(34.928684, 33.601897)	Chalkidikis 14, Aradippou 7101, Cyprus
3	2032	(34.837709, 32.433364)	Miltiadi Stylianou Ave 4, Tala 8577, Cyprus
4	5200	(35.040622, 33.974079)	2XRF+6J Paralimni, Cyprus
5	1216	(35.107277, 33.211556)	4646+WJ Ayioi Trimithias, Cyprus

Figure 2 Sample details of the service points proposed by the developed model

For each of the customers in the final dataset, we have calculated whether they are closer to a service point of CompA, or CompB, with the majority of customers being closer to a service point of CompA, as shown in Figure 4.

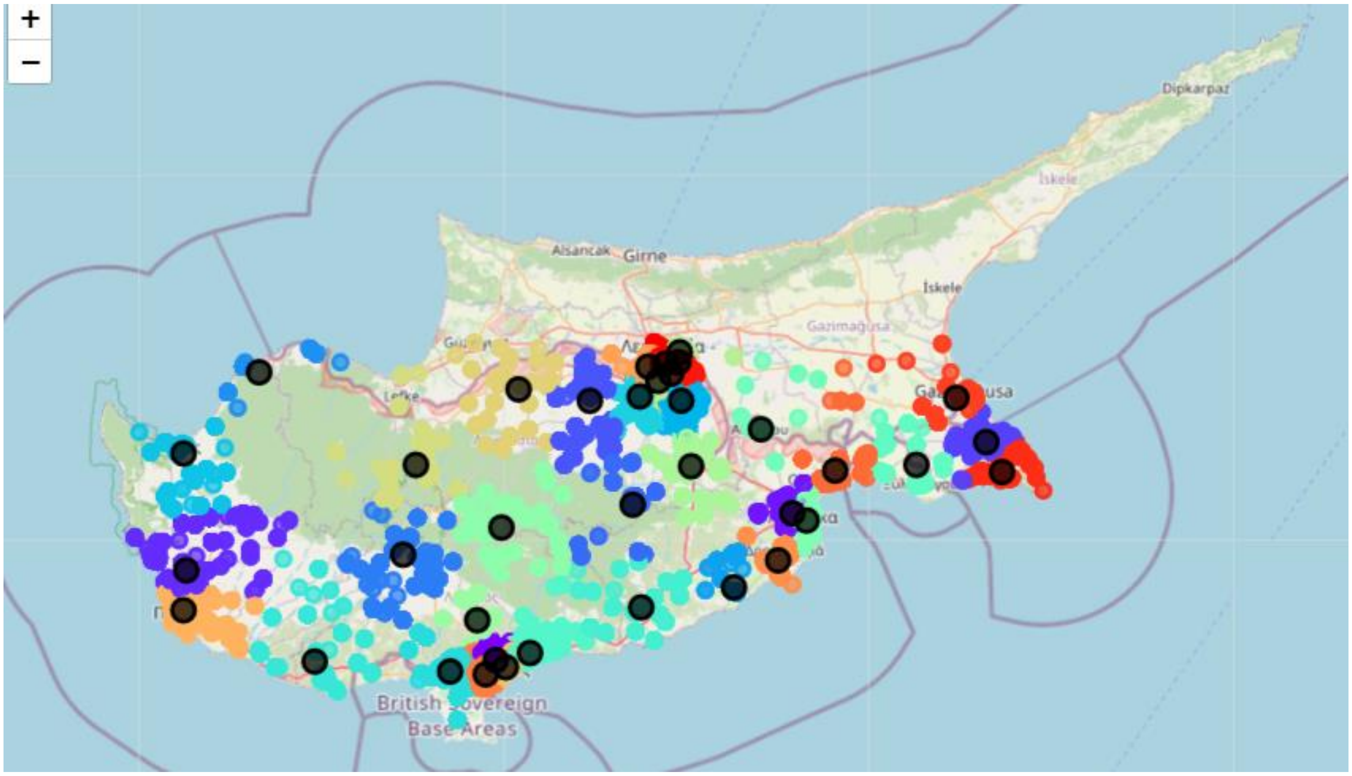

Figure 3 A colour map visualization of the 37 clusters returned as results of the proposed k-Means clustering model

Another interesting result was the boxplot of the distance distribution of the 3 instances shown in Figure 5. The plot showed that even though there are a lot of distance outliers in all 3 instances, yet the model's outliers are significantly less than CompA and CompB, and the maximum distance of the model is at 20km, compared to the max distance of CompA and CompB, being 27km and 35km respectively. This suggest that adopting the model's service point suggestions, that the farthest away customer, would have to travel 20 km to a service point, which is an improvement for both companies.

This was even further investigated by finding the distance to the nearest service point at the 95th percentile. The 95th percentile is an important percentile in that it measures the maximum distance to be travelled to the nearest service point by 95% of the sample population of customers. The top 5% could be dismissed as outliers. The measurements are shown in Figure 6, and are significant, because it reveals that through the optimal service point placement, the distance to the nearest service point was almost halved from around 7km for both CompA and CompB, to 3,2km for the model's service point suggestions. It is also important to mention that this was achieved without significantly distorting the minimum distance to travel at the 5th percentile – the minimum distance that 95% of the population would have to travel – does not exceed the value of 0.2km.

The distance to the nearest service point was also measured by the distance range to the nearest service point for comparison purposes, shown in Figure 7. From the plot we notice that CompA's service points are placed to serve the vast majority of its customers under <1km (160k). The number of customers being serviced from service points in the 1-5km range are only 25% of the customers serviced in <1km (40k), and the number continues to decrease further for the 5-10, 10-20 and >20 km range. CompB's service points follow a more balanced pattern, closer to the model's pattern.

The proposed model's service points, suggest service point placement such that the number of customers being serviced by service points in the <1km and 1-5km range do not present great variability as CompA's service point placement. It is also noticeable that there are no customers in the >20km range.

Another interesting measurement was the count of the number of customers per service point. The customers were counted to the nearest service point in each of the 3 instances for comparison. We have then proceeded to define the branch size as a function of the number of customers in its vicinity as shown by Figure 8. The results are shown in Figure 9.

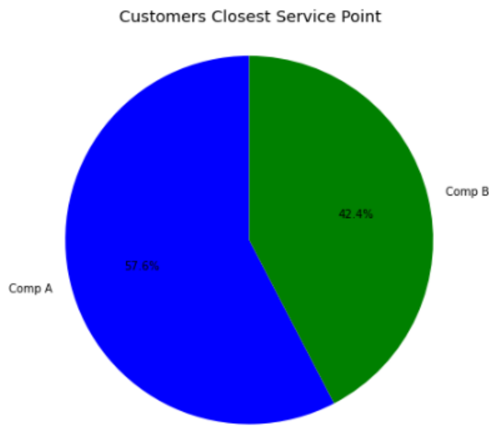

Figure 4 Percentage of customers to nearest service point by company

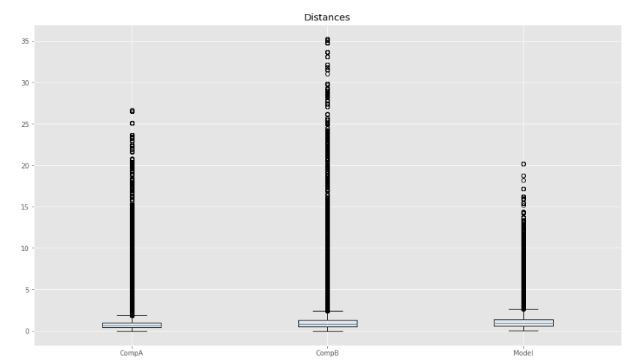

Figure 5 Boxplot of the distance to the nearest service point for each company and the model's proposal

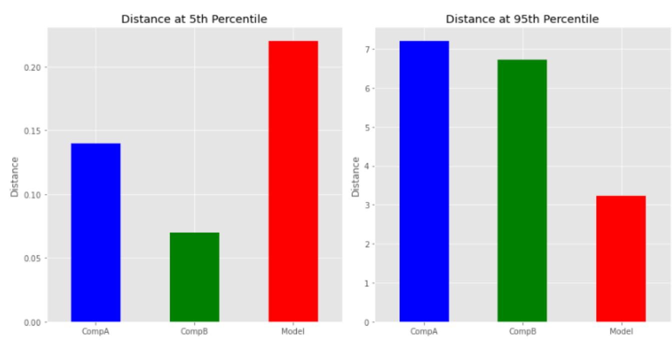

Figure 6 Distance in km to the nearest service point at 5th and 95th percentile for each company and the model's proposal

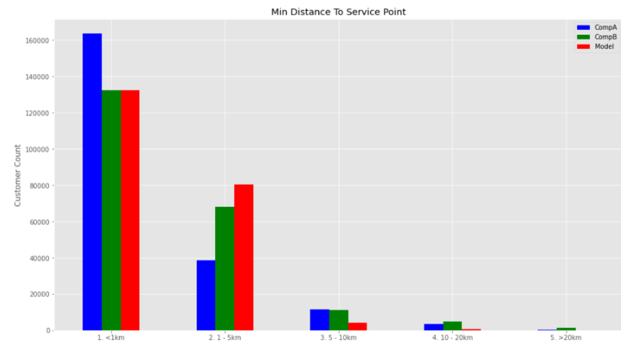

Figure 7 Minimum distance range to the nearest service point

	ServicePointSize	CustomersServiced	CompA	CompB	Model
0	ATM	<500	2	0	4
1	Small	500-2.000	22	8	12
2	Medium	2.000-10.000	32	30	16
3	Large	10.000-20.000	3	3	3
4	ExtraLarge	>20.000	1	1	2

Figure 8 Definition of Service Point Size as a function of the number of customers in its vicinity

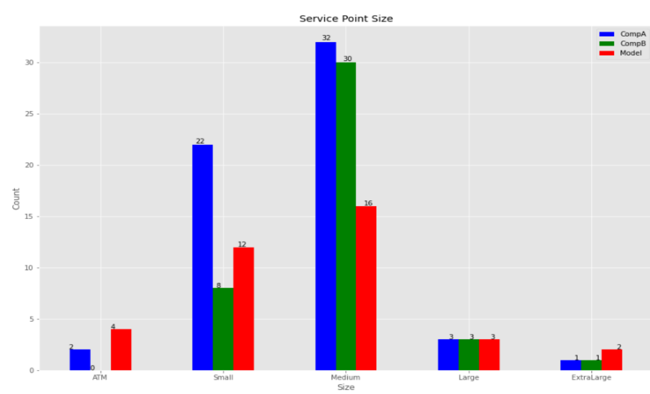

Figure 9 Number of branches by service point size for each company and the model's proposal

From the distribution plot in Figure 9, we notice that CompA and CompB have mostly small and medium-sized branches in terms of the customers being serviced. The model proposes a wider and more balanced range of branch sizes. This is mostly noticeable in the "Medium" category of branches, where Company A and Company B have 32 and 30 branches respectively, whereas the model suggests only 16 medium sized branches.

The size of the branch can also be interpreted in terms of the staff needed to service the branch's customers. The reduced number of branches, as well as the calculated staff needs, translate to reduced operating costs, while maintaining a high quality of service to the company's customers.

A sample differentiation of the service point placement, depicted by the difference in colors, and the service point size, depicted by the marker's radius, is shown in Figure 10, for a sample area in Nicosia city.

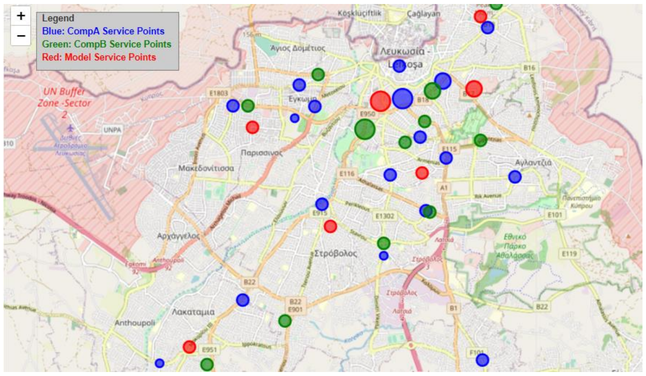

Figure 10 Sample service point distribution and service point size of the model's suggested service points in Nicosia

CONCLUSIONS

For the purposes of this model development, we have succeeded in proposing a near-optimal service point placement for a retailer company, by using the company's customers' geolocation as a data source. This data-driven approach, and utilization of machine learning method, can benefit a retailer company whose interest is to minimize distance to its customers. The suggested service point locations appear strategically spaced-out, and the results show that we have succeeded in that 95% of the retailer's customers will be serviced by service points within less than 3.2km. At the same time, we have proposed fewer service points, with calculated branch sizes in terms of staff needed to serve customers in the service point's vicinity, that would give the retailer company the advantage of reduced operating costs while offering better customer service through data-driven insight on the branch size.

FUTURE WORK

Future work on this subject would be to identify different options to customer geolocation gathering. One such option would be through cellular phone data.

Another option would be to compare the actual service point preferences of customers, versus the proposed / nearest service points as identified in the model, to understand if any other factors affect the customer's preferred service point location, other than its proximity.

ACKNOWLEDGMENTS

I am deeply grateful to my supervisor, Dr. Demetris Trihinas, for the constant encouragement, invaluable guidance and support throughout the course of this project. Their expertise, knowledge, and dedication have been instrumental in shaping this work and fostering my growth as a data scientist. I am truly fortunate to have had the opportunity to benefit from his mentorship.

REFERENCES

[1] Miliotis P, Dimopoulou M, and Giannikos I. A hierarchical location model for locating bank branches in a competitive environment. International transactions in operational research 2002; 9: 549-565

[2] Min H, Melachrinoudis E. The three-hierarchical location-allocation of banking facilities with risk and uncertainty. International transactions in operational research 2001; 8: 381-401

[3] Church R, ReVelle C. The maximal covering location problem. Papers in regional science 1974; 32: 101-118

[4] Church RL, Roberts KL. Generalized coverage models and public facility location. 1983; 53: 117-135

[5] Karasakal O, Karasakal EK. A maximal covering location model in the presence of partial coverage. Computers & Operations Research 2004; 31: 1515

[6] Eiselt HA, Marianov V. Gradual location set covering with service quality. Socio-economic planning sciences 2009; 43: 121-130

[7] Alexandris G, Giannikos I. A new model for maximal coverage exploiting GIS capabilities. European journal of operational research 2010; 202: 328-338

[8] CY Opendata Portal. National opendata portal. Retrieved 12/2022, from https://www.data.gov.cy/?language=en

[9] CY Opendata Portal, Τμήμα Εφόρου Εταιρειών και Διανοητικής Ιδιοκτησίας. Μητρώο εγγεγραμμένων εταιρειών, εμπορικών επωνυμιών και συνεταιρισμών στην κύπρο. Retrieved 12/2022, from https://www.data.gov.cy/node/4016?language=en

[10] Wikipedia. Haversine formula. 2023. Retrieved 15/04/2023, from https://en.wikipedia.org/wiki/Haversine_formula

[11] Wikipedia. Euclidean distance. 2023. Retrieved 05/05/2023, from https://en.wikipedia.org/wiki/Euclidean_distance

[12] Wikipedia. K-means clustering. 2023. Retrieved May 5,2023, from https://en.wikipedia.org/wiki/K-means_clustering

[13] Wikipedia. Determining the number of clusters in a data set. 2023. Retrieved 15/04/2023, from https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set

[14] Wikipedia. Elbow method (clustering). 2023. Retrieved May 5,2023, from https://en.wikipedia.org/wiki/Elbow_method_(clustering)