Capstone Project: Clustering Medical Centers in Puebla, Mexico

Gabriela A. Rodríguez Ruiz

May 14, 2019

I. Introduction

In 2017, the Mexican Association of Faculties and Schools of Medicine registered 38,600 applications for medical residences, of which only 21% obtained a place [1]. There are many reasons because not all of the medical applicants get a place in a hospital, the most obviously is because there are not enough places. However, in some cases, the applicants get a place in the hospital but them don´t want to take it because of the location of the hospital. In these cases, the doctors prefer to continue doing a social service in hospitals and wait one more year to apply to residences again. Generally, at each application, the doctors can establish three hospitals in which, if they be accepted, they are assigned one of the three hospitals established before.

Considering it and using the Foursquare data location [2], in this work we are interested in clustering the medical assistance centers of the Puebla state in Mexico, taking into account the amenities that are around the medical centers, in order to doctors can choose similar medical centers as their three options.

II. Data

In order to achieve the goal of this work, that is clustering the medical assistance centers based in the amenities around them, we are going to use the dataset of Puebla, Mexico Locations of Municipal Buildings, Schools, Parks, Medical Assistance Centers, and Gardens. This dataset is provided by NYU ,Spatial Data Repository [3] page and it contains the locations of the medical assistance centers, parks, schools, temples, municipal buildings and so on. Specifically, for this work we're interested in the locations of the medical assistance centers in Puebla, Mexico. The dataset contains 30340 instances and for each there are their latitude and longitude coordinates. An example of the dataset can see in Figure 1.

Moreover, in combination with the dataset describe before, we are going to use the foursquare data location, in order to obtain the amenities/venues around the each medical assistance center in Puebla, Mexico.

97	Centro de Asistencia Médica	19.050759	-97.943619
98	Instalación Deportiva o Recreativa	19.054576	-97.940450
99	Templo	19.047924	-97.942135
100	Centro de Asistencia Médica	19.048078	-97.942903
101	Escuela	19.060329	-97.992656
102	Pozo	19.061705	-97.994583
103	Escuela	19.059127	-97.998996
104	Centro de Asistencia Médica	19.053554	-97.995795

Figure 1. Example of Dataset of locations of medical centers in Puebla, Mexico.

III. Methodology

The general scheme of the proposed method in this work can be shown in Figure 2.

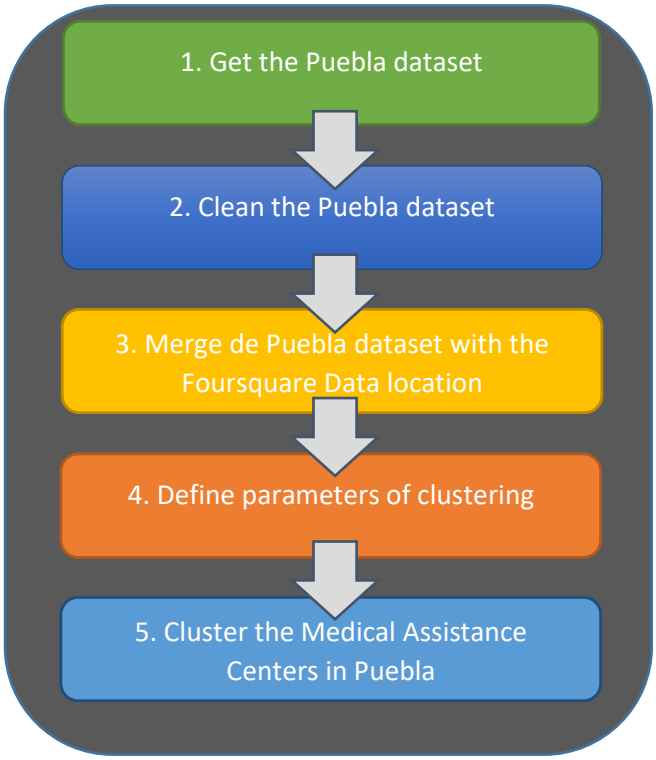


Figure 2. Proposed Methodology for clustering Medical Assistance Centers in Puebla, Mexico.

As a first stage of the methodology, we get the "Puebla" dataset from the Spatial Data Repository [2]. This dataset, contains latitude and longitude coordinates of different services, schools, parks, temples, gardens, recreation areas and medical assistance centers from Puebla State in Mexico. In total, there are 30340 instances, from which 2816 are Medical Assistance Centers (See Figure 3).

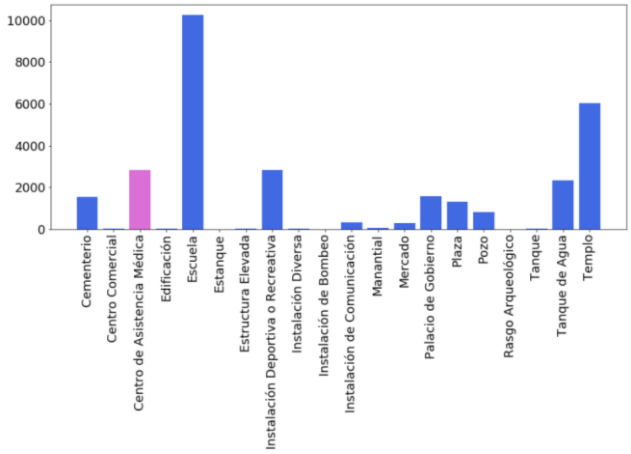


Figure 3. Instances from Medical Assistance Centers.

At the second stage of the methodology, we clean the "Puebla" Dataset, keeping only with that rows that belong to the Medical Assistance Centers class. Due to the number of medical centers in

all the State are huge, we define a radius with difference on 0.2 in latitude and longitude from the center of Puebla City and we only keep with 283 medical centers (See Figure 4).

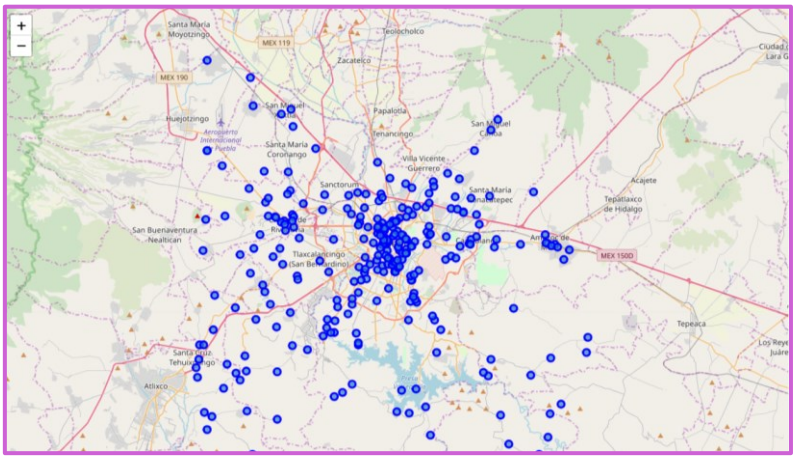


Figure 4. Reduction to only 283 Instances from Medical Assistance Centers.

Once we have the latitude and longitude coordinates of the Medical Centers in Puebla, we merged the data with the data obtained by Foursquare Data Location. The first step was establish a connection to Foursquare by means of the Client_ID, the Client_Secret and then by means of the explore command, we get the near venues to each medical center.

Then at the four step of the methodology proposed in this work, we define the parameters of clustering, in this case, we plot the elbow method in order to see how many clusters are convenient to take. The plot is not so clear, but there is a soft elbow taking 3 clusters; thus, for clustering the medical centers que took 3 clusters (see Figure 5).

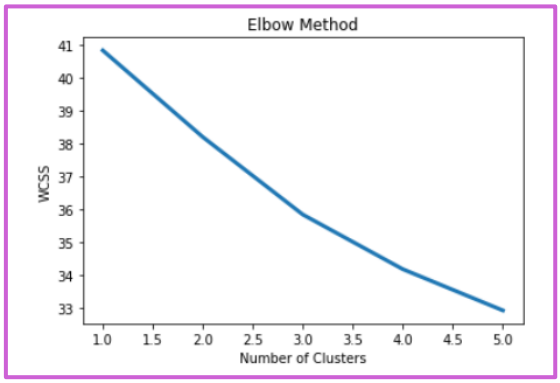


Figure 5. Selection of clusters by Elbow Method.

After defining the number of clusters, the final step of our methodology is clustering the medical centers. For this purpose, we applied the K-Means Method to our dataset in order to cluster the medical centers. The aggrupation is based in the similarities of the medical centers taking into account the 10 near venues of each medical center. In the next section, the results obtained by our methodology will be analyzed.

IV. Results

As we mentioned before, we tested out methodology with the "Puebla" dataset in conjunction with the Foursquare Data Location. After cleaning our dataset, we obtained 283 instances for medical assistance centers. For each one of these medical centers, que obtained 10 venues from Foursquare. After applying the K-Means Clustering, we clustered in three groups the medical centers (See Figure 6).

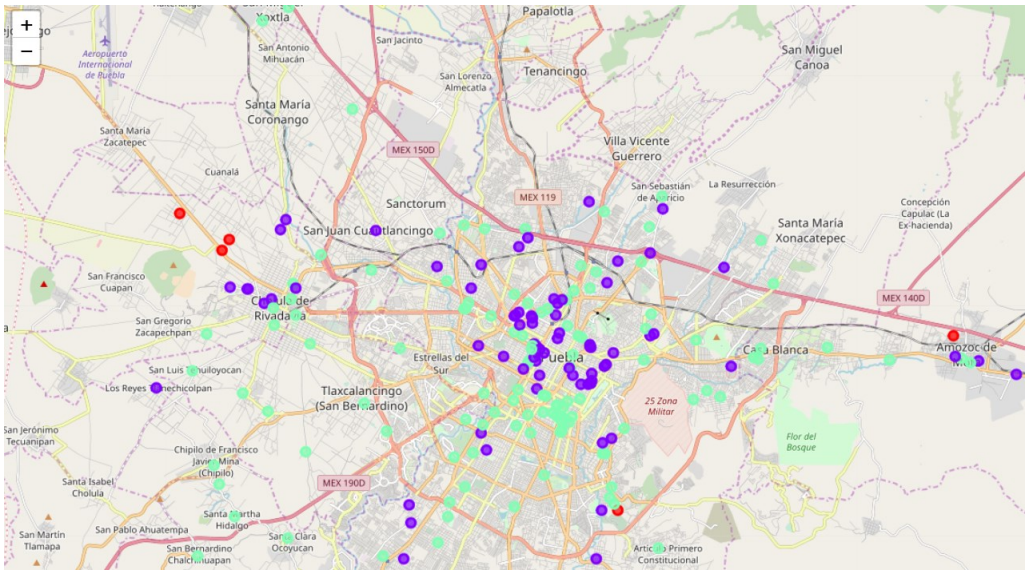


Figure 6.Clustering Medical Assistance Centers in Puebla, Mexico.

In order to analyze each cluster and understand which type of venues are more common for each cluster, we calculated the mode value for each of the 10 most commons venues. And the results are shown in Figure 7, 8 and 9 respectively.

We can see, that cluster one is mainly composed by zoo exhibit, event space and Flower shops. Also, cluster 2 is composed by Taco places, Pizza places and Food and Drink Shops. Finally, cluster 3 is mainly composed by Taco Place, Convenience Store and Food Court.

2	3	4	5	6	7	8	9	10
Zoo Exhibit	Event Space	Zoo Exhibit	Exhibit	Flower Shop	Flea Market	Fish Market	Fish & Chips Shop	Fast Food Restaurant

Figure 7. Most common venues for Cluster 1.

2	3	4	5	6	7	8	9	10
Taco Place	Taco Place	Pizza Place	Food & Drink Shop	Food	Food	Flower Shop	Fish Market	Fish & Chips Shop

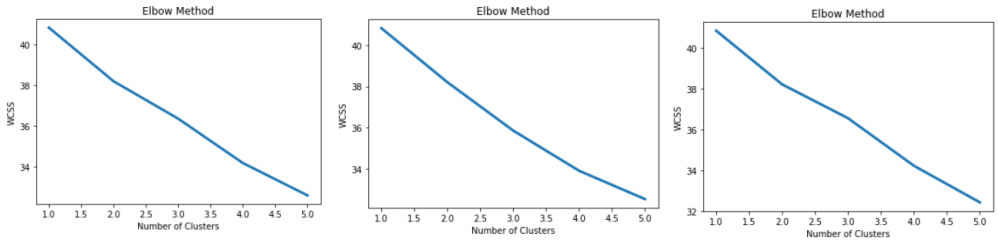
Figure 8. Most common venues for Cluster 2.

2	3	4	5	6	7	8	9	10
Taco Place	Convenience Store	Food Court	Food & Drink Shop	Food	Flower Shop	Flea Market	Fish Market	Fish & Chips Shop

Figure 9. Most common venues for Cluster 3.

V. Discussion

As we saw in last section, the results were very interesting about the most common venues around each medical center cluster. However, about the selection of the number of clusters to take into the K-Means Method, the plot of the Elbow Method wasn't clear. I think that if we use more instances and more venues it could help to identify more precisely the right number of clusters to use in K-Means method. Also, in the process of getting the right number of clusters, we tried with different values for max_iter and random_state in the KMeans method of the Sklearn library.



a) b) c)

Figure 10. Different values for random_state in KMeans method a) 2, b) 3 and c) 5

VI. Conclusion

In this study, I analyzed the similitude of Medical Assistance Centers of Puebla Mexico, based in the common near venues of each medical center. I identified that the common venues for each cluster are different and they can be a guide for the doctors at the moment of choose their three options for medical residences allocations. This work is only a first step as the study of similitudes of different medical centers. As a future work, it could be a good idea to explore and analyzed the medical centers of different states in Mexico, because the assignation of medical residence is a national level.

VII. References

1. Mejía, Ximena (06/2018). Aumenta demanda de estudiantes de Medicina en México. Excelsior. https://www.excelsior.com.mx/nacional/aumenta-demanda-de-estudiantes-de-medicina-en-mexico/1245232
2. Foursquare Developers. (2019). NYC, SF and Chicago. https://es.foursquare.com/developers/apps
3. Instituto Nacional de Estadística y Geografía México, (2015). Puebla, Mexico Locations of Municipal Buildings, Schools, Parks, and Gardens. NYU, Spatial Data Repository https://geo.nyu.edu/catalog/nyu-2451-37011