Clustering Neighborhoods in Mexico City

Uriel Ibarra

February 29, 2020

1. Introduction

1.1. Background

After nearly 700 years of history, Mexico City is one of the largest and most important human metropolises on the planet and the oldest city in the Americas. Its history is surely complex, and despite the country's issues, its capital is a beautiful place with many things to offer. It has thousands of attractions, and its streets and buildings shows the colorful history of a society constructed through centuries of hard work. But CDMX, as known in Spanish, is also an evolutionary place, and it cannot be otherwise if it is the largest urban center in the country. This capital had been lagging behind and far away of the smartness that involves other metropolis, but recently its new government has implemented policies to boost citizen well-being and deliver more efficient, sustainable and inclusive urban services and environments as part of a collaborative, multi-stakeholder process. Technology will help us to understand what has worked, what has not worked, and what can be improved. Digitalization has reach the oldest city and has come to stay.

Today more than ever, access to information is a right, data has become an indispensable tool for the improvement of human activities. As a small contribution to this trend, I will try to make an analysis of the most common factors stakeholders would take into account to select a place to move and thus group the neighborhoods into clusters.

Photo: santiagoarau.com

1.2. Problem

Data that might contribute to find an optimal location to live in include the Human Development Index (HDI) of the area, number and distance to nearby transport stations, health centers, recreation centers and food shops and also, in the case of Mexico City, the seismic risk level of the area. This project aims to group neighborhoods with others that share similar aspects based on these data.

1.3. Interest

People who are looking for a new place to move in Mexico City would be very interested in knowing the characteristics of the different neighborhoods of the city and which ones resemble each other.

2. Data acquisition and cleaning

2.1. Data sources

Mexico City's new Digital Agency of Public Innovation will be the source of the geospatial, IDH, transport, health centers and seismic data, while Foursquare API will provide the data of the food shops and recreation centers scattered across the city:

- Neighborhood location
- Neighborhood Human Development Index
- Seismic Risk Atlas
- Public transport stations (Metro, Bus, Trolley)
- Public health center location
- Recreation center location (*Art and entertainment* Foursquare category)
- Food shop location (*Food and drink shop* Foursquare category)

All these datasets, however, lack data in certain aspects. For example, the neighborhood location dataset misses the location of four neighborhoods, while the HDI one misses the information of about five hundred. Also, the seismic dataset contains all the risk locations in the city which don't correspond, under no circumstances, to the neighborhood coordinates. To fix all these issues, different data science techniques were used. Kepler tool was helpful to generate different map types.

Additionally, the neighborhoods were clustered by density in order to reduce the universe of locations to be analyzed. One of the resulted clusters was chosen arbitrarily before gathering recreation centers and food shops data because of Foursquare's free membership limitations.

2.2. Data cleaning

- Neighborhoods

This dataset originally contained several geospatial and administrative data of 1812 neighborhoods from Mexico City. There were four missing location values, whereby the corresponding neighborhoods were discarded.

- Neighborhood Human Development Index

This dataset originally contained all data used to compute the HDI of 1447 neighborhoods in Mexico City and the HDI itself. The missing data were replaced with the mean of the neighborhood indexes of the corresponding borough.

- Seismic risk, public transport stations and health centers

These datasets were used in their original form.

3. Explanatory data analysis and feature selection

- Neighborhoods

Neighborhoods ID, name, latitude and longitude were extracted from this dataset. A visualization of the data is shown in Fig. 1.

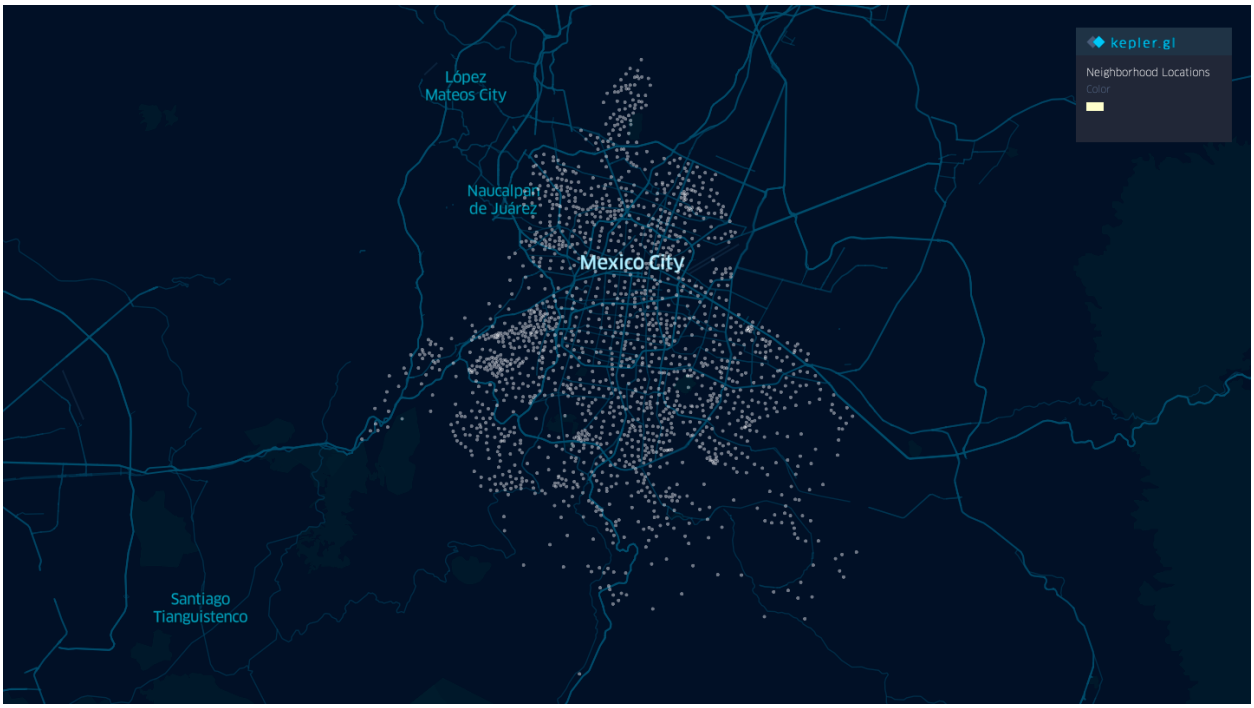

Fig. 1 Mexico City neighborhoods.

- Neighborhood Human Development Index

Neighborhood ID and HDI were extracted from this dataset. Later, it was merged with the geospatial data to be displayed. A visualization of the data is shown in Fig. 2.

- Seismic risk

Since this dataset originally contained the data about the Mexico City seismic locations (latitude and longitude) and their risk level, a classification algorithm were used to find the risk level of every neighborhood. A visualization of the original data is shown in Fig. 3.

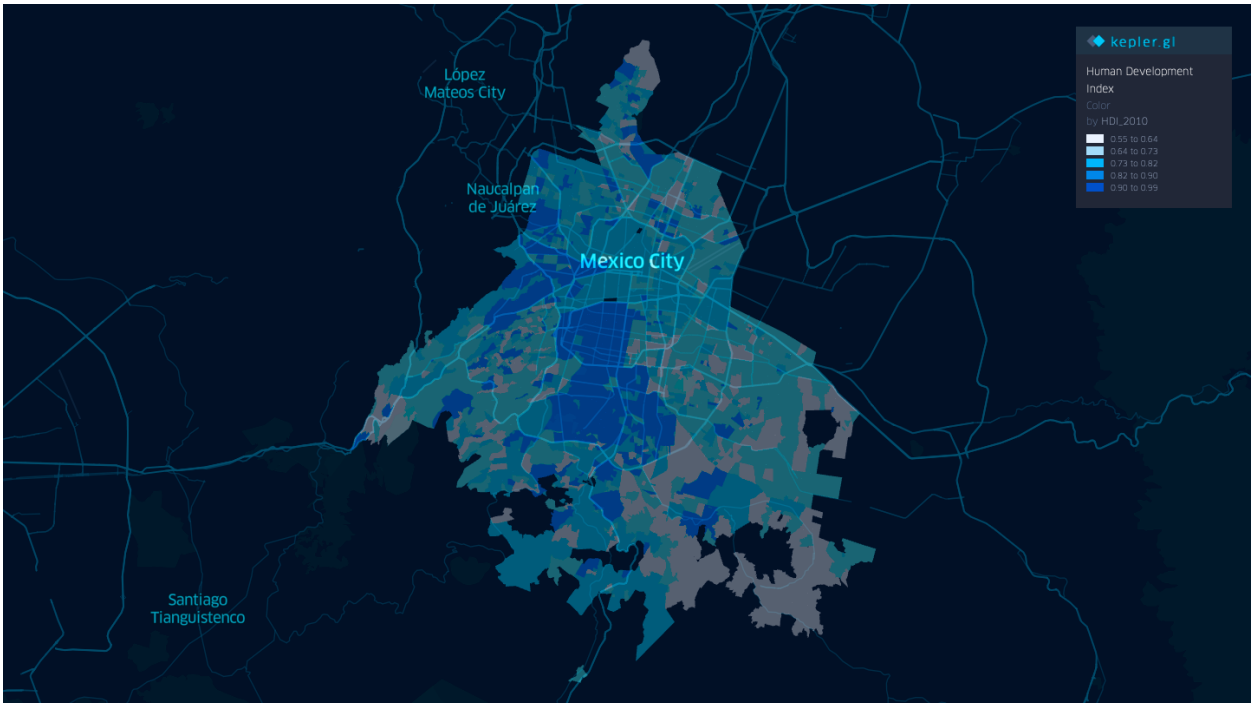

Fig. 2 Mexico City neighborhoods by Human Development Index.

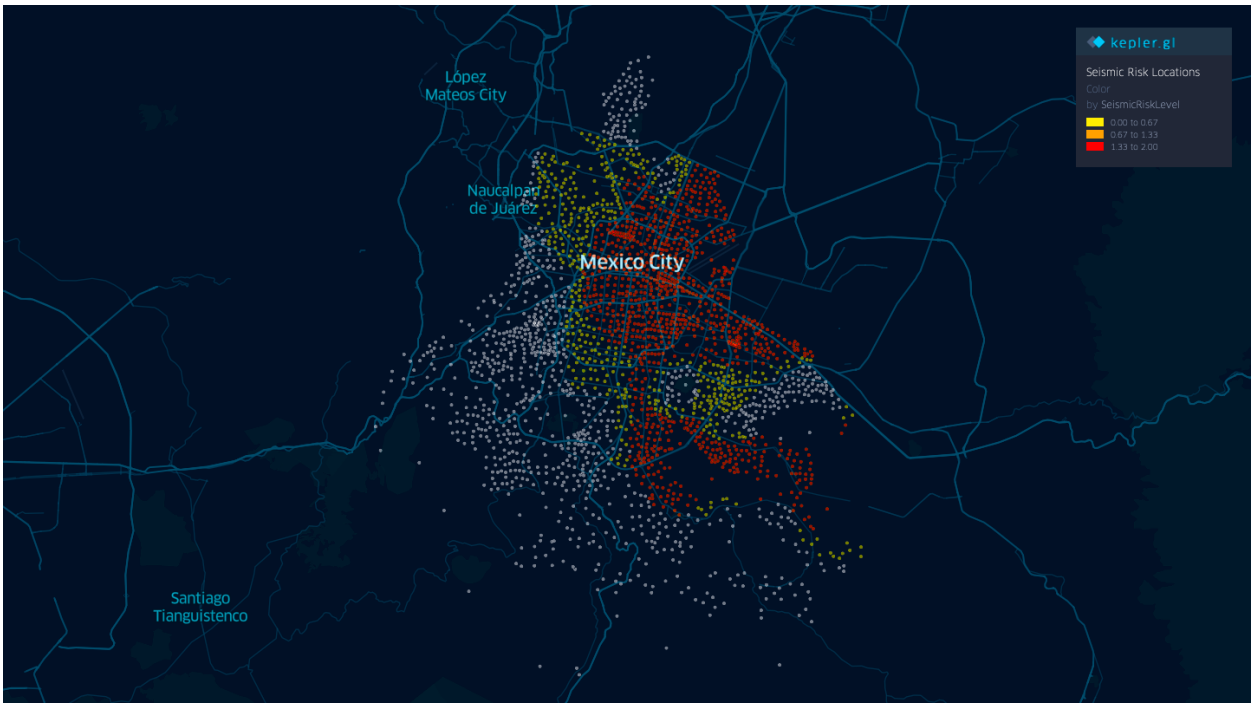

Fig. 3 Mexico City seismic risk locations.

K-nearest neighbors algorithm was used to classify each Mexico City neighborhood according to its proximity to seismic risk locations. The final value of K was determined by training the algorithm for different values of K and selecting the most accurate (Fig. 4).

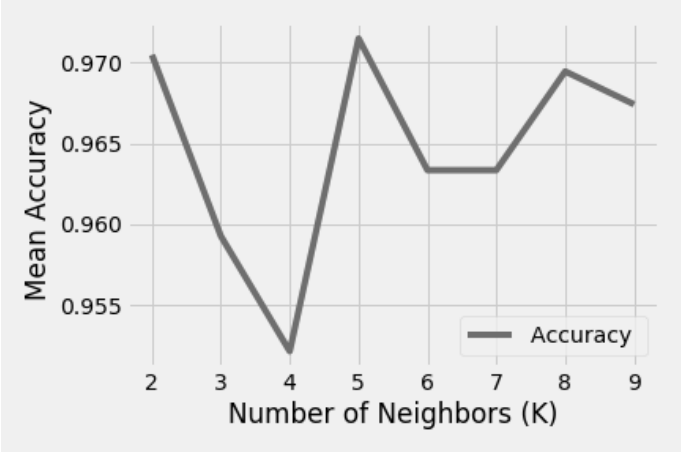

Fig. 4 Accuracy for different K values for the neighborhoods classification by seismic risk (K = 5).

Then, the model obtained was used to label every Mexico City neighborhood. A visualization of the result is shown in Fig. 5.

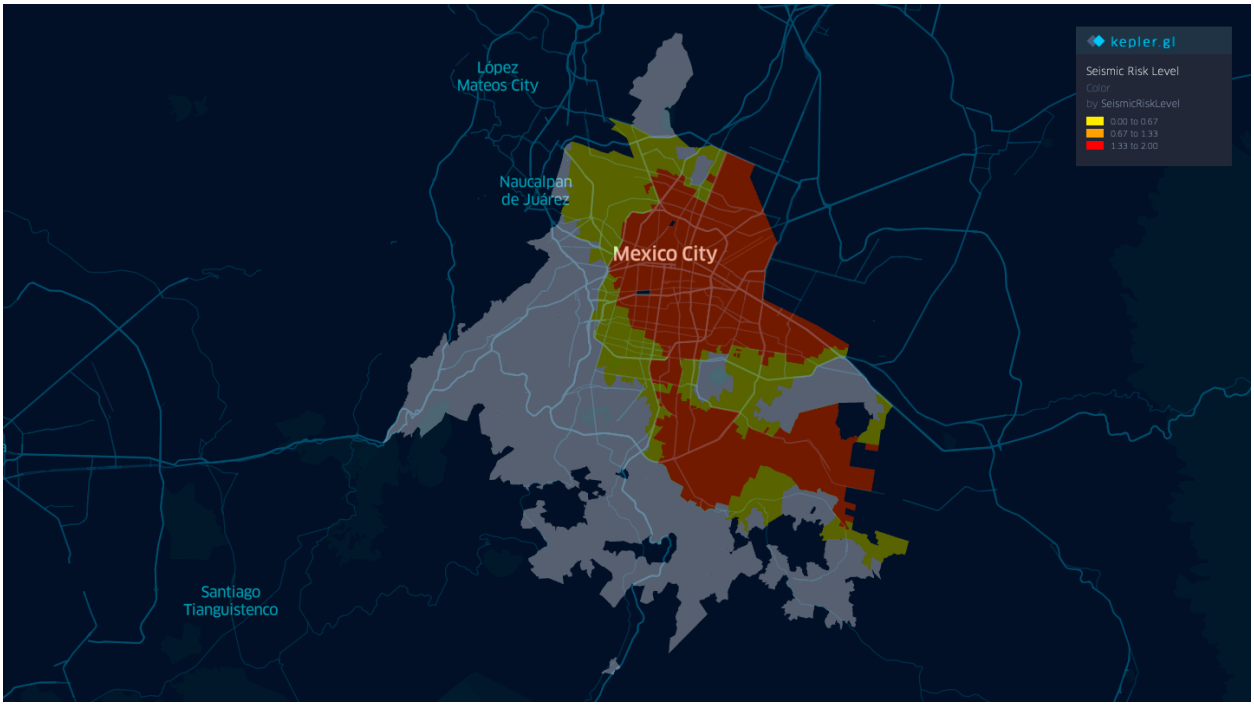

Fig. 5 Mexico City neighborhoods by seismic risk level.

- Public transport stations and health centers

Public transport stations and health centers name, latitude and longitude were extracted from this dataset. A visualization of the data is shown in Fig. 6. The coordinates were used to count the number of stations and centers within 1 and 5 km around each neighborhood, respectively, and also the distance to the nearest transport station and health center.

Fig. 6 Mexico City public transport stations and health centers.

- Recreation centers and food shops

Before gathering the venues data, neighborhoods were clustered by density using K-means algorithm. Elbow method was used to determine the number of clusters (Fig. 7).

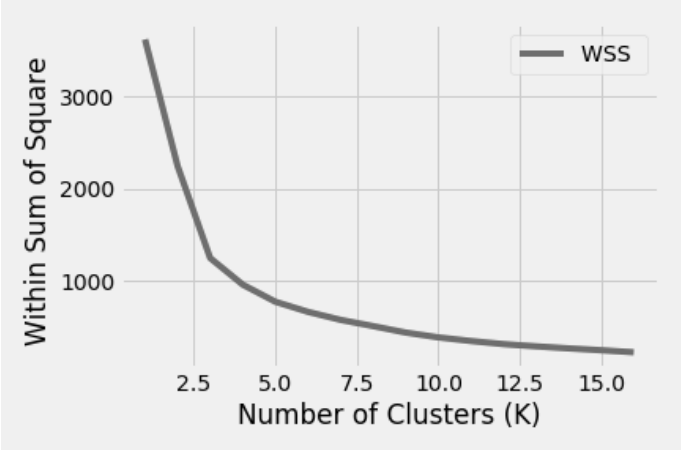

Fig. 7 Elbow method for the neighborhoods clustering by density (K = 4).

Then, the model obtained was used to label every Mexico City neighborhood (Fig. 8).

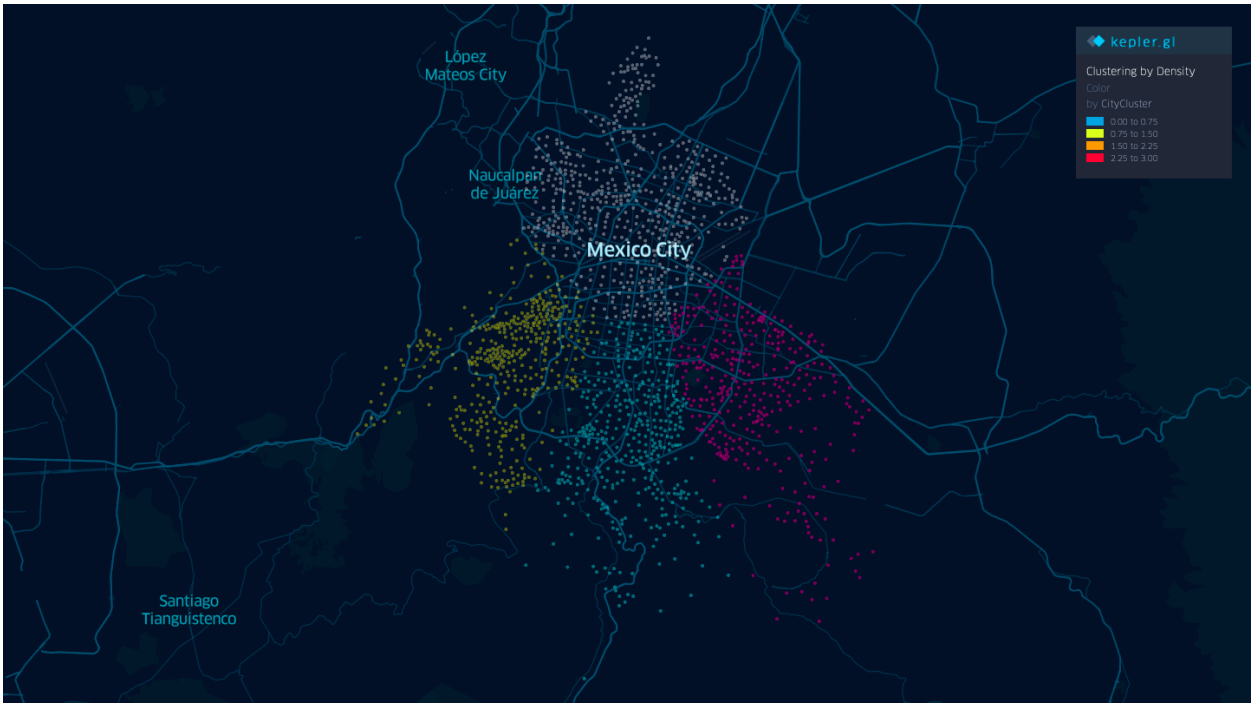

Fig. 8 Mexico City neighborhoods clustered by density.

Then, one of the clusters was selected for the final analysis. Arbitrarily, the zone which contains the well-known Mexico City Main Square: The Zocalo (cluster 1), was chosen (Fig. 9).

Fig. 9 Mexico City selected cluster.

This cluster contains 583 neighborhoods located in the Norther Central part of the city. The locations of the center of every of them were used to finally gather the venues data from the Foursquare API, which were filtered using the categories provided by the company. The latitude and longitude of venues belonging to the *Art and entertainment* and *Food and drink shops* categories were put into a dataset. These coordinates were used to count the number of centers and shops within 3 and 1 km around each neighborhood, respectively, and also the distance to the nearest recreation center and food shop.

4. Methodology

4.1. Target

The next features will be used to group the Northern Central part of Mexico City into clusters that share similar aspects to live in.

- The human development index.
- The seismic risk level.
- The number of transport stations within 1 km around.
- The distance to the nearest transport station (trolley, bus or metro).
- The number of health centers within 5 km around.
- The distance to the nearest public health center.
- The number of recreation centers within 3 km around.
- The distance to the nearest recreation center.
- The number of food shops within 1 km around.
- The distance to the nearest food shop.

4.2. Clustering model

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm was used to cluster neighborhoods by confront, according to parameters mentioned above. The final value of epsilon, which defines the maximum distance between two points, was determined by calculating the distance to the nearest *n* points for each point, sorting and plotting the results (Fig. 10). Then, the optimal value for epsilon was found at the point of maximum curvature.

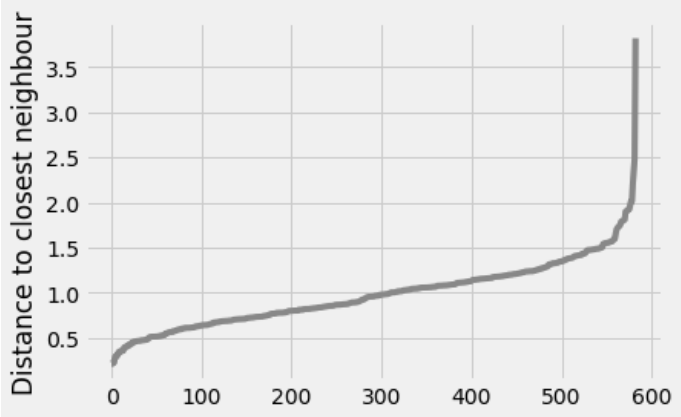

Fig. 10 Knee method for the neighborhoods clustering by comfort factors (eps = 1.6).

5. Results

5.1. Solution

The model obtained was used to label every Northern Central Mexico City neighborhood. A visualization of the result is shown in Fig. 11. Candidate neighborhoods were grouped into three clusters, labeled with numbers 0, 1 and 2. DBSCAN algorithm is sensitive to noise, thus noisy neighborhoods were labeled with "-1".

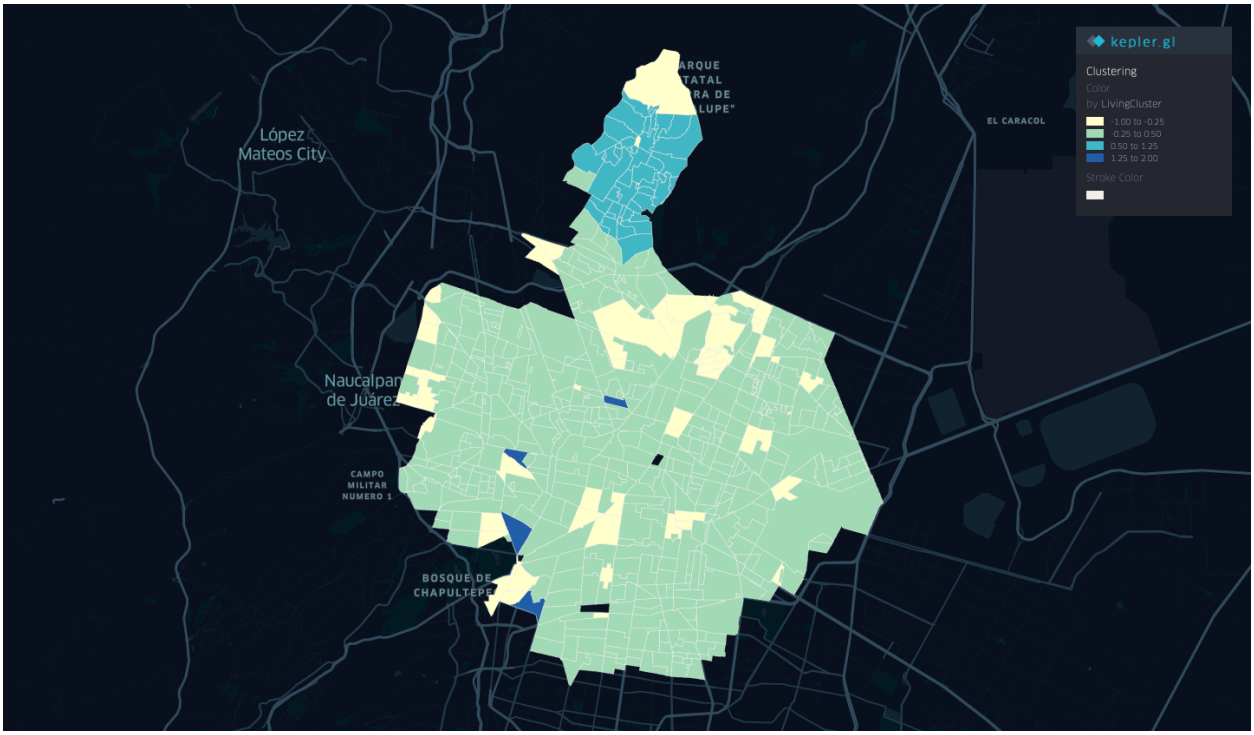

Fig. 11 Northern Central Mexico City neighborhoods clustered by confort.

6. Discussion

6.1. Clustering distributions

The next figures show the side-by-side histograms per each factor considered in the analysis. The left side of each figure shows the general distribution of neighborhoods, while the right one shows the clustered distribution colored. As we can see, cluster 0 represents most of the neighborhoods, as its distributions emulate the general one. Cluster 1 describes a specific part of the city where seismic risk is low compared with the majority and where transport stations, health centers, food shops, and recreation centers are far away. Finally, cluster 2 groups only four neighborhoods of the city that stand out among the rest. These are *Defensores de la República*, *Popotla*, *Hipódromo* and *Anzures*, which have a HDI score bigger than the average, suitable seismic risk level and a good number of nearby transport stations, health centers, food shops, and recreation centers.

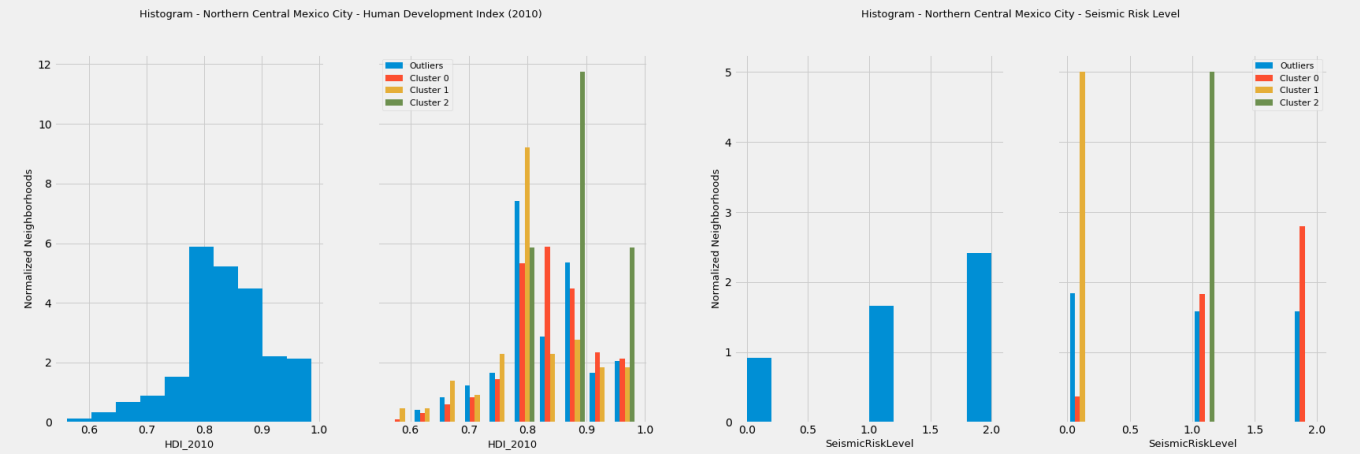

Fig. 12 Northern Central Mexico City neighborhoods HDI score and seismic risk level distributions.

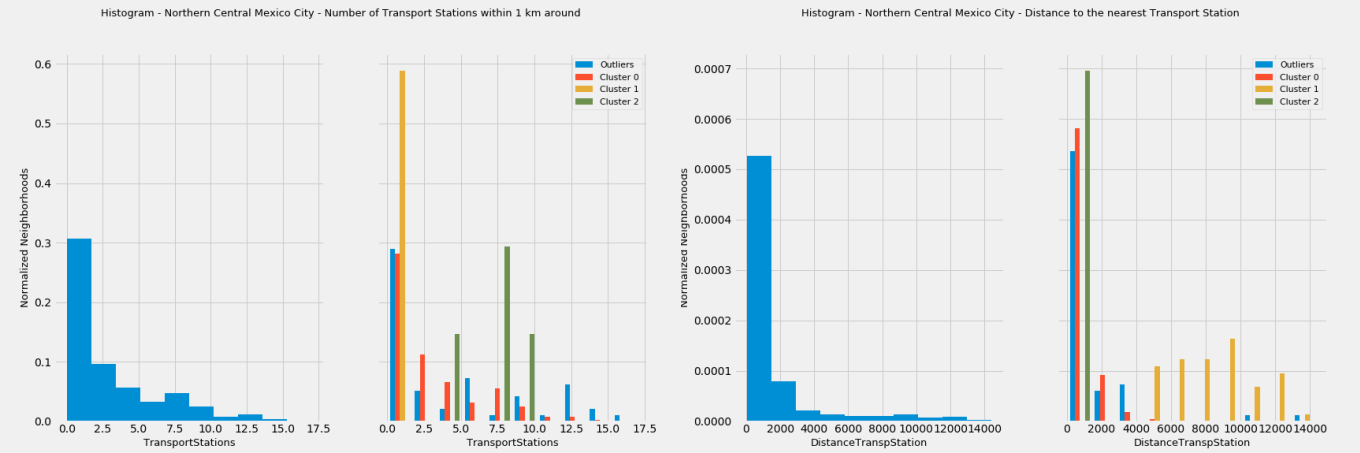

Fig. 13 Northern Central Mexico City neighborhoods public transport stations availability distributions.

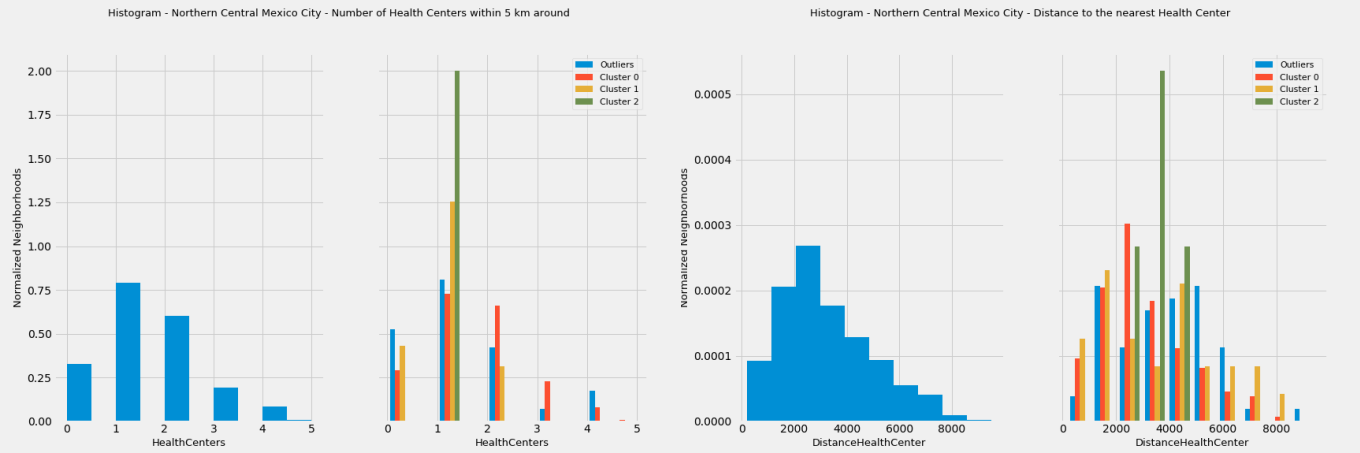

Fig. 14 Northern Central Mexico City neighborhoods public health centers availability distributions.

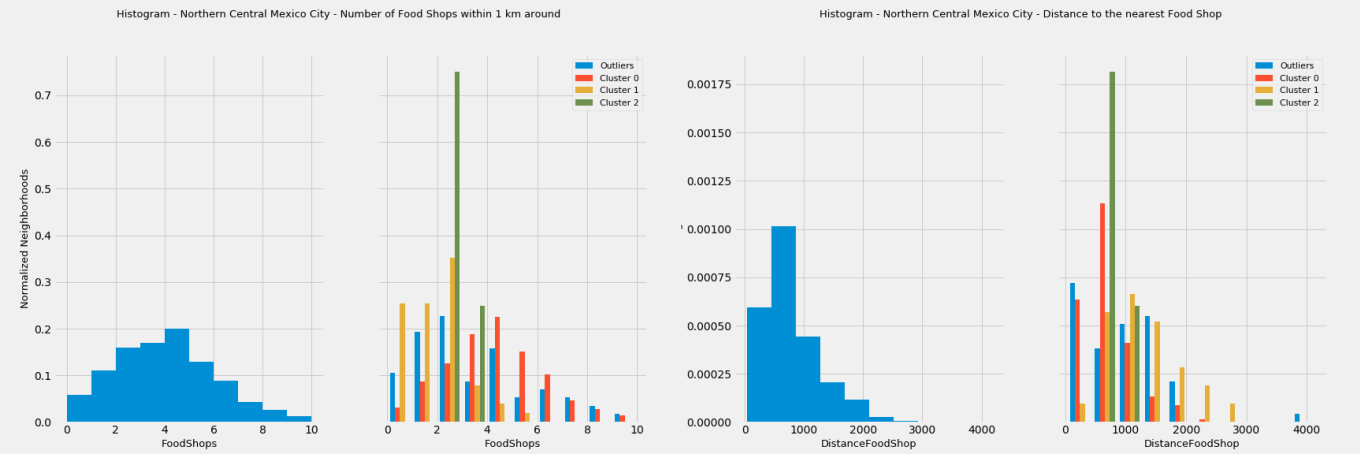

Fig. 15 Northern Central Mexico City neighborhoods food shops availability distributions.

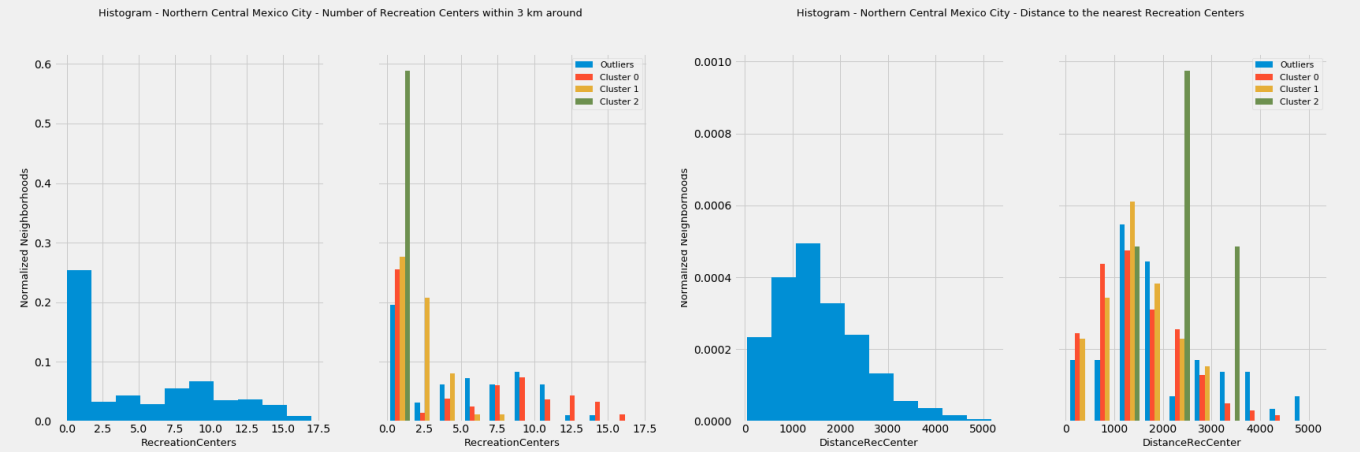

Fig. 16 Northern Central Mexico City neighborhoods recreation centers availability distributions.

7. Conclusion

The purpose of this project was to identify the optimal places to live in Mexico City according to factors that could influence people to move to a particular area. Clustering of neighborhoods was then performed in order to create a tool to sustain that kind of decisions and insights that were starting points for exploration by stakeholders.

The final decision on optimal house location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood, etc.