Geographical cluster of the possible COVID-19 infection points in México City

Andres J. Ramos

1. Introduction

1.1 Background

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a virus that had it's initial outbreak in Wuhan, China in December 2019 and has stopped most of the economic and social activities around the world. Seven months after the first outbreak was recorded, some governments have decided to restart global economic activities gradually to resume them with great caution. Given that currently social gathering points could be the main sources of contagion to initiate a resurgence. It is important to know which places and establishments will be most frequented based on social behaviors.

1.2 The problem

In Mexico City, for example, there are boroughs that are presenting large outbreaks and want to return to economic and social activities as soon as possible. Therefore, it would be very useful for health authorities to obtain a classification of the places of social events near the geographical areas that have registered the most infected people, in order to take preventive measures when social coexistence are reactivated.

This project aims to classify geographical locations in Mexico City that could be considered a future focus of infection from highest to lowest risk. Through the use of the Foursquare API that would provide information on the inflow into establishments around the infected people, implementing the unsupervised K-Means algorithm and choropleth map.

1.3 Interest

This information can be useful when health authorities in Mexico City need to establish prevention measures to avoid contagion in public places, as well as to determine which types of meeting places can be opened and which should remain closed until contagion decreases.

2. Data acquisition and cleaning

2.1 Data source

For this project, the data from Mexico City are obtained from the official URL of the government of México whose author is “Secretaria de Salud de la Ciudad de México”:

https://datos.cdmx.gob.mx/explore/dataset/casos-asociados-a-covid-19/table/?disjunctive.resultado

This dataset provides all cases related to COVID-19 up to 6/2/2020 and is updated regularly every week.

2.2 Important keywords

Due to the increasing rate of infections, health services are insufficient, so the Secretary of Health has decided to classify the cases as *outpatient* and *inpatient.*

- **Outpatient** cases (“Casos Ambulatorios”): refers to non-severe SARs-CoV-2 positive cases where the person is not hospitalized and is asked to undergo quarantine at home.
- **Inpatient** (“Casos Hospitalizados”): Refers to severe SARs-CoV-2 positive cases requiring urgent medical care and hospital isolation.

Due to misinformation, apathy and lack of interest from most of the mexican population regarding the COVID-19 situation, it represent a serious risk of virus resurgence. Very much *outpatients* are confusing the symptoms of SARs-CoV-2 with the influenza and don't take proper isolation measures.

2.3 Data cleaning and feature selection

The original data set includes not only the positive and negative cases, but the geographical location, clinical conditions of the patient, place of birth, place of hospital where they were treated, registration ID, etc.

The data cleaning aims to create a clean Dataframe that contains:

- Only Positive cases
- Only outpatient cases

First, all the columns that indicate if the patient have a chronic condition has been dropped, this includes: EPOC, diabetes, asthma, hypertension, obesity, etc. Also, the columns that

contain information like birth site, nationality, hospital sector, update date, gender and home state.

Second, the rows that contain negative cases, hospitalized cases, and deceased were dropped.

Third, the result dataset only contains the columns “borough” and “result”.

Kept features	Drop features	Reason for dropping features
Borough	‘DIABETES','EPOC','ASMA','INMUNOSUPRESION', 'HIPERTENSION','OTRA COMPLICACION', 'CARDIOVASCULAR', 'OBESIDAD', 'RENAL CRONICA', 'TABAQUISMO', 'OTRO CASO', 'MIGRANTE', 'PAIS NACIONALIDAD', 'PAIS ORIGEN', 'UNIDAD DE CUIDADOS INTENSIVOS', 'RANGO EDAD','ID_REGISTRO', 'num_fallecidos', 'num_hospitalizados', 'EDAD'	These features do not provide relevant information.
Results	'FECHA ACTUALIZACION','FECHA SINTOMAS','ORIGEN','SECTOR', 'ENTIDAD UNIDAD MEDICA','SEXO', 'ENTIDAD NACIMIENTO', 'ENTIDAD RESIDENCIA', 'INTUBADO','NEUMONIA', 'NACIONALIDAD', 'EMBARAZO','HABLA LENGUA INDIGENA',	

The final result is ordered as following:

	Borough	Results
0	Azcapotzalco	1366
1	Benito Juárez	933
2	Coyoacán	1647
3	Cuajimalpa de Morelos	588
4	Cuauhtémoc	1398
5	Gustavo A. Madero	3069
6	Iztacalco	1366
7	Iztapalapa	4259
8	La Magdalena Contreras	678
9	Miguel Hidalgo	1028
10	Milpa Alta	970
11	Tlalpan	1965
12	Tláhuac	1376
13	Venustiano Carranza	1253
14	Xochimilco	1978
15	Álvaro Obregón	1561

Figure 1 Data set head

2.4 Obtaining geographical data

Using the Foursquare API in order to get the most popular places, it is necessary to assign the coordinates to the dataframe.

	Borough	Results	Latitude	Longitude
0	Azcapotzalco	1366	19.485815	-99.184206
1	Benito Juárez	933	19.380470	-99.163243
2	Coyoacán	1647	19.328040	-99.151063
3	Cuajimalpa de Morelos	588	19.318707	-99.323203
4	Cuauhtémoc	1398	19.441613	-99.151864

Figure 2 Dataframe with coordinates head

The next step is retrieving each venue in a range of 3000 meters and limit them to 10 places. This data frame contains 152 places, and on an average 10 venues per neighborhood. The result data set is the following:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Azcapotzalco	19.485815	-99.184206	La Conchería CDMX	19.483789	-99.185843	Bakery
1	Azcapotzalco	19.485815	-99.184206	Café ONCE28	19.484427	-99.185720	Breakfast Spot
2	Azcapotzalco	19.485815	-99.184206	La Perla Tapatía	19.483741	-99.185856	Mexican Restaurant
3	Azcapotzalco	19.485815	-99.184206	Neko Café	19.484152	-99.183326	Japanese Restaurant
4	Azcapotzalco	19.485815	-99.184206	Centro Verde Azcapotzalco	19.487757	-99.182125	Garden

Figure 3 Venues per neighborhood

In order to get the top 10 places of each venue per borough, it is necessary to encode these categories. This is going to be helpful to estimate the frequency of each venue.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Azcapotzalco	Mexican Restaurant	Breakfast Spot	Japanese Restaurant	Art Gallery	Bakery	Gym / Fitness Center	Garden	Garden Center	Fair	Farm
1	Benito Juárez	Ice Cream Shop	Pet Store	Bakery	Gourmet Shop	IT Services	Japanese Restaurant	Park	Yoga Studio	Sushi Restaurant	Sporting Goods Shop
2	Coyoacán	Ice Cream Shop	Coffee Shop	Art Museum	Restaurant	Food Court	BBQ Joint	Pizza Place	Taco Place	Gym / Fitness Center	Farm
3	Cuajimalpa de Morelos	Mexican Restaurant	Mountain	Historic Site	Track	Park	Restaurant	Farm	Scenic Lookout	Event Space	Factory
4	Cuauhtémoc	Scenic Lookout	Historic Site	Taco Place	Sushi Restaurant	Hostel	Art Museum	General Entertainment	Public Art	Museum	Monument / Landmark
5	Gustavo A. Madero	Burger Joint	Mexican Restaurant	Ice Cream Shop	Taco Place	Restaurant	Sports Club	Market	Flower Shop	Food Court	Event Space
6	Iztacalco	Racetrack	Music Venue	Sushi Restaurant	Stadium	Sporting Goods Shop	Sporting Event	Baseball Stadium	Yoga Studio	Food Court	Donut Shop
7	Iztapalapa	Taco Place	Mexican Restaurant	Beer Garden	BBQ Joint	Food Truck	Burger Joint	Garden Center	Factory	Fair	Farm
8	La Magdalena Contreras	Theme Park	Playground	Soccer Stadium	Nature Preserve	Plaza	Rock Climbing Spot	Memorial Site	Soccer Field	Paintball Field	General Entertainment
9	Miguel Hidalgo	Ice Cream Shop	Park	Cocktail Bar	Cycle Studio	Department Store	Hotel	Jewelry Store	Men's Store	Yoga Studio	Sporting Event
10	Milpa Alta	Factory	Camera Store	Yoga Studio	Garden	Event Space	Fair	Farm	Flower Shop	Food Court	Food Truck
11	Tlalpan	Coffee Shop	Ice Cream Shop	Bar	Italian Restaurant	Mexican Restaurant	Event Space	Dessert Shop	Arcade	Bakery	Flower Shop
12	Tláhuac	Mexican Restaurant	Seafood Restaurant	Taco Place	Fair	Lounge	Café	Pizza Place	Food Court	Donut Shop	Event Space
13	Venustiano Carranza	Airport Lounge	Spanish Restaurant	Donut Shop	Coffee Shop	Sandwich Place	Seafood Restaurant	Snack Place	Hotel	Steakhouse	Airport Service
14	Xochimilco	BBQ Joint	Flower Shop	History Museum	Coffee Shop	Candy Store	Mexican Restaurant	Beer Garden	Garden Center	Yoga Studio	Food Truck
15	Álvaro Obregón	Taco Place	Brewery	Paintball Field	National Park	Golf Driving Range	Farm	Café	Garden	Seafood Restaurant	Yoga Studio

A quick analysis of the table reveals that the top 10 is mostly composed of fast food restaurants, followed by recreation centers and restaurants in general.

3. Classification model

Classification is an approach to supervised machine learning that can be interpreted as a way to categorize or classify some unknown elements in a discrete set of classes. For this case the information to be classified is the common venues in México city.

The model used is *K-means*, a method that aims to partition *n observations* into *k-clusters*, with each *observation* belongs to the cluster with the nearest mean (cluster centers or cluster centroid).

K-mean classified the information into two clusters as following:

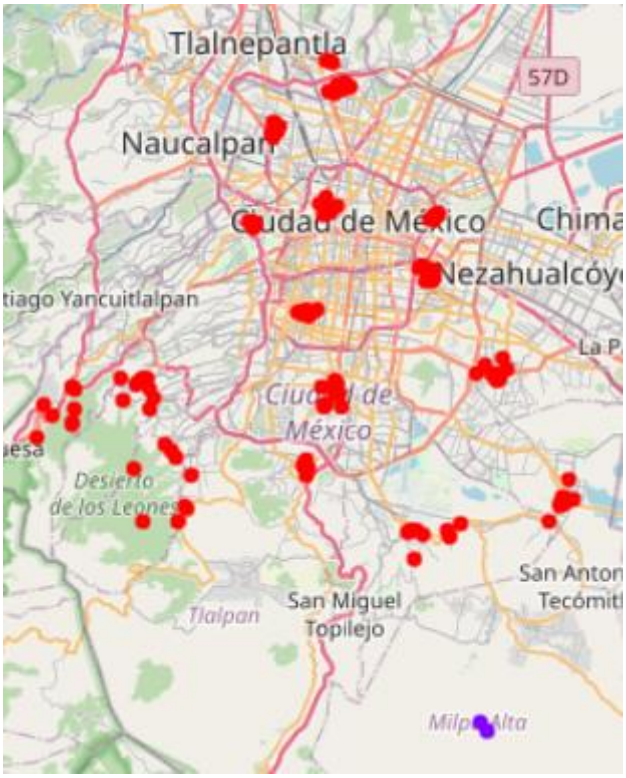

Figure 4 Cluster Map

This may indicate that there is a great similarity between the locations in Mexico City or that the model may have underfitting.

4. Plot the data into a map

Generate a choropleth map is the following step. In order to achieve this, I used the first dataframe of coronavirus and a geojson file obtained from the official page of México City:

https://datos.cdmx.gob.mx/explore/dataset/alcaldias/table/

The choropleth map focuses on showing the territorial divisions and concentration of COVID-19 cases.

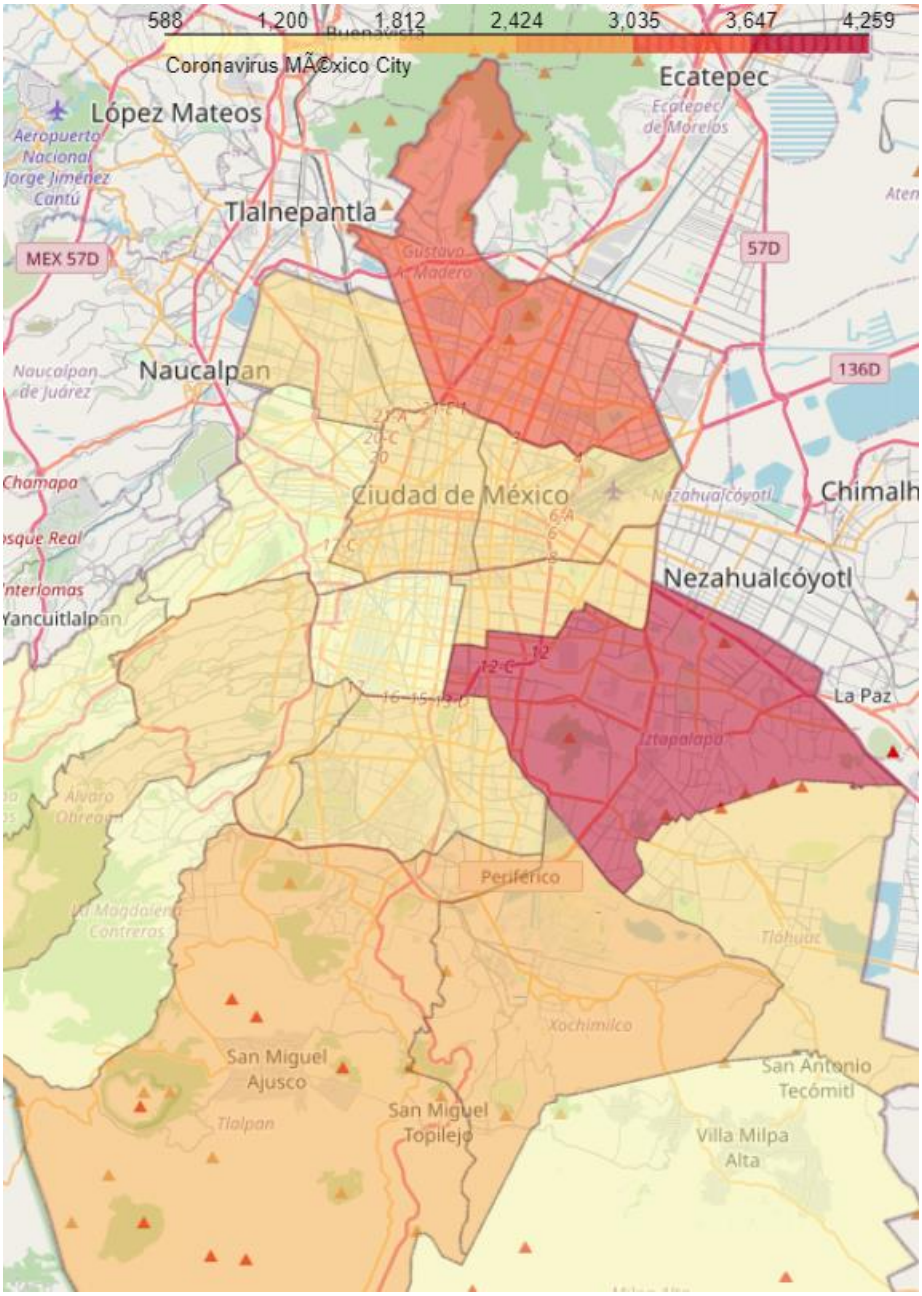

Figure 5 Choropleth Map - Cases per Neighborhood

The last step is mapping the top venues around the map and combine the choropleth map.

The result is a map which shows the public places most susceptible to generate infections.

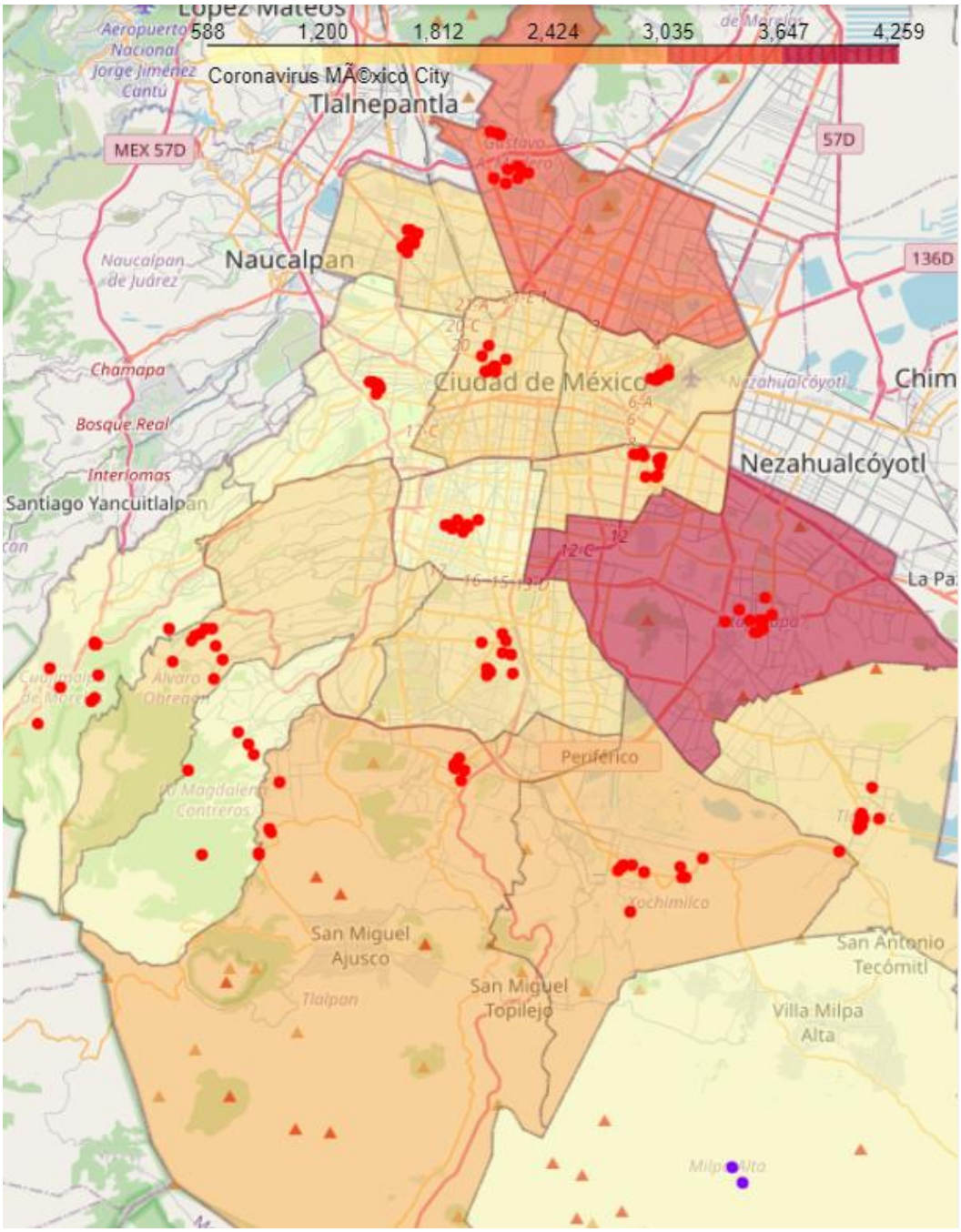

Figure 6 Map of possible places of infection

5. Conclusions

In this study, I analyzed the possible places that could cause coronavirus outbreaks in México City based on the most popular places in México. Maybe because of the limited foursquare (free) account, some important places were left out, this map can give an idea of the behavior of people in each suburb. This information could be useful to create better health measures to prevent a resurgence of the virus.

6. Future directions

I think that in the future I could improve the accuracy if I classify the places by the most visited areas or maybe get what places are around the infected people. In order to implement these updates, it is necessary to find the datasets containing this information.