Machine learning against pollution: clustering analysis of Greater Mexico City's air quality

A. Payen-Sandoval

1 Introduction

Greater Mexico City (GMX), which comprises Mexico City and the east half of the surrounding State of Mexico, is one of the largest, most populated regions in the world, housing over 20 million people[1]. In 1992, GMX was considered the most polluted urban area in the world[2]. It got much better, for some years, but around 2016, pollution notoriously started coming back[3]. Since then, multiple measures have been implemented, from strict regulation of vehicle traffic to outright recommending people not to go out during certain times of day, when pollution reaches its peak values. These measures are not enough, however, and pollution is taking its toll on public health and the economy.

GMX is an interesting candidate to study pollution: it contains heavily urbanized areas, rural areas, mountains, lakes, basins and, of course, an immense amount of buildings. The goal of this work is to analyze the influence that different types and amounts of venues can have on pollution. This and future studies of the same type could be used by governments to measure the impact of building new venues before they exist, and regulate accordingly, to avoid future environmental contingencies.

2 Data

The air quality accross all of GMX is being measured by the government in 38 fixed monitoring stations (figure 1). The coordinates, names and codes of each station were obtained from the webpage for air quality monitoring of Mexico City[4] and arranged as a dataframe (figure 2).

The Foursquare API was used to obtain the types and amounts of venues around each station (figure 3) in a 2 km radius.

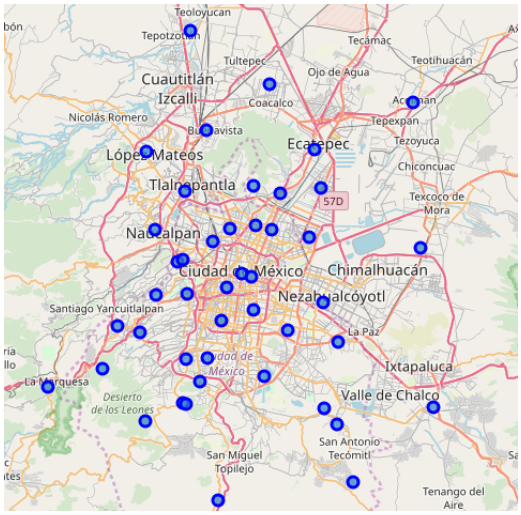

Figure 1: Distribution of stations around GMX. Each blue dot is a station.

	Name	Latitude	Longitude	Altitude
ACO	Acolman	19.635501	-98.912003	2198
AJU	Ajusco	19.154674	-99.162459	2953
AJM	Ajusco Medio	19.272100	-99.207658	2619
ATI	Atizapán	19.576963	-99.254133	2341
BJU	Benito Juarez	19.371612	-99.158969	2250

Figure 2: Some of the stations. The index of the table is the code that identifies each station. Their names, coordinates and altitudes are listed.

	Stores	Restaurants & Bars	Medical	Buildings & Offices	Hotels & Hostels	Arts	Museums & Galleries	Plazas	Farms	Schools	Gyms & Sports	Stadiums & Theaters	Bus Stations	Fields
AJM	7	28	0	1	0	0	0	0	1	0	4	0	0	0
ATI	4	22	0	1	0	0	0	0	2	0	4	0	0	2
BJU	48	54	0	3	1	2	0	0	4	0	12	7	0	3
CCA	10	55	0	0	0	2	1	0	2	2	9	7	0	1
CUA	26	41	0	2	0	0	0	0	0	0	15	0	0	2

Figure 3: Venues around some of the stations.

There are six pollutants of interest and for which data is readily available at [4]:

- Carbon monoxide (CO), measured in parts per million (*ppm*);
- sulfur dioxide (SO_2), measured in parts per billion (*ppb*);

- ozone (O_3), measured in *ppb*;
- nitrogen dioxide (NO_2), measured in *ppb*;
- particulate matter less than 10 microns in diameter (PM_{10}), measured in $\mu g/m^3$; and
- particulate matter less than 2.5 microns in diameter ($PM_{2.5}$), measured in $\mu g/m^3$.

Each pollutant has an associated csv file, which consists of hourly averages of the measured pollutants, for each day of the year, for all 38 stations (figure 4).

	FECHA	HORA	ACO	AJM	ATI	BJU	CAM	CCA	CHO	CUA	...	SAG	SFE	SJA	TAH	TLA	TLI	UAX	UIZ	VIF	XAL
0	2019-01-01	1	-99	1	5	2	3	2	1	2	...	11	2	-99	1	4	5	2	5	7	4
1	2019-01-01	2	-99	1	5	2	3	2	1	2	...	11	2	-99	1	4	6	2	5	8	4
2	2019-01-01	3	-99	1	5	2	3	2	1	2	...	12	2	-99	1	5	6	2	5	8	4
3	2019-01-01	4	-99	1	5	2	3	2	1	2	...	12	2	-99	1	5	6	2	5	8	4
4	2019-01-01	5	-99	1	5	2	3	2	1	2	...	12	2	-99	1	5	6	2	6	8	4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8755	2019-12-31	20	-99	4	8	6	8	4	-99	2	...	-99	2	-99	2	20	36	4	6	28	-99
8756	2019-12-31	21	-99	4	10	6	10	4	-99	2	...	-99	2	-99	2	22	36	4	6	28	-99
8757	2019-12-31	22	-99	4	10	6	10	4	-99	4	...	-99	2	-99	2	22	36	4	6	28	-99
8758	2019-12-31	23	-99	4	12	6	10	4	-99	4	...	-99	2	-99	2	22	36	4	6	28	-99
8759	2019-12-31	24	-99	4	12	6	10	4	-99	4	...	-99	2	-99	4	24	36	4	6	28	-99

Figure 4: Raw data extracted from the SO_2 file. The first column, 'FECHA', is the date. The second column is the hour at which the average concentration of SO_2 was taken, and the remaining columns correspond to the codes of each station (see figure 2). Each -99 value represents missing data.

Correlations will be made between yearly average pollution values and venue data. From these correlations, appropriate variables will be chosen to form clusters of stations, and correlations will be determined, again, for each separate cluster. It is expected, given the environmental diversity of GMX, that each cluster will present a different set of correlations, and thus, provide clues about the underlying mechanisms of pollution in each separate cluster. Once these mechanisms are spotted, we will incorporate basic physics and chemistry into the discussion.

References

[1] Borbet, T. C., Gladson, L. A., & Cromar, K. R. (2018). Assessing air quality index awareness and use in Mexico City. BMC public health, 18(1), 538. doi:10.1186/s12889-018-5418-5

[2] S. Campbell, Monica (12 May 2016). "Why Mexico City's bad air can't be ignored – or easily fixed". PRI's The World. Retrieved 5 February 2020.

[3] Schachar, Natalie. "Mexico City Tries New Tactics Against an Old Enemy: Smog". CityLab. Retrieved 5 February 2020.

[4] SEDEMA. Gobierno de la Ciudad de México. Data available at *http://www.aire.cdmx.gob.mx/default.php*