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 to stay indoors 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, forests, lakes, basins and, of course, an immense amount and diversity of human-made venues. In GMX, The IMECA ('metropolitan index for air quality', by its spanish meaning) is a standarized unit used to measure the concentrations of multiple pollutants in the air. Being a dimensionless unit, it provides easy understanding of its meaning to the general population, and allows direct comparison between concentrations of pollutants, regardless of their molecular masses and chemical properties. IMECA scores of 0 to 50 are considered good, from 50 to 100 they are considered regular, and from 100 to 150 they are considered bad. More than 150 is considered very bad, and more than 200 is extremely bad. Environmental contingencies in GMX are announced when any pollutant has an IMECA score of 150 or more.

The goal of this work is to analyze the influence that different types and amounts of venues can have on the IMECA scores of carbon monoxide (CO), ozone (O_3), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), particulate matter with less than 10 micrometers in diameter (PM10) and particulate matter with less than 2.5 micrometers in diameter (PM25), in the vicinity of air quality measuring stations spread across GMX, and to infer sets of rules regarding venues and their effects on pollutants. Doing this requires that we group stations with similar rules together, for which a clustering algorithm is ideal.

This and future studies of the same type could be used by governments and private institutions to measure the impact of building or adapting new venues and avoid future environmental contingencies, or to estimate maintenance costs due to pollution.

2 Data

The air quality across all of GMX is measured at all times 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).

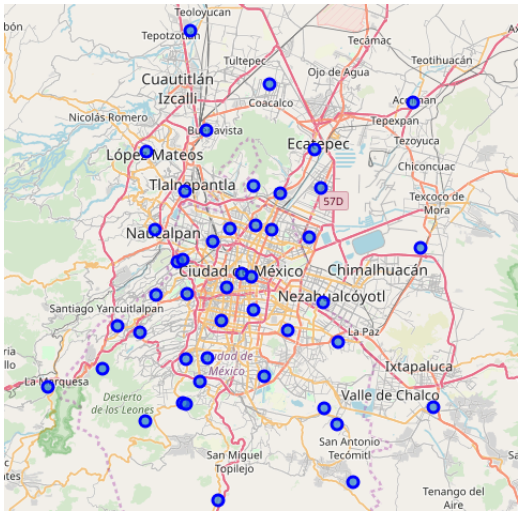

Figure 1: Distribution of stations around GMX. Each blue dot is a station.

	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. Each station is represented in figure 1.

There are six pollutants of interest and for which data is readily available at [4]: CO, measured in parts per million (ppm); SO_2, measured in parts per billion (ppb); O_3, measured in ppb; NO_2, measured in ppb; PM10, measured in $\mu g/m^3$; and PM25, also measured in $\mu g/m^3$. Each pollutant has an associated csv file, which consists of hourly averages of these pollutants, for each day of the year, for all 38 stations (figure 3). Each pollutant file was loaded onto a separate dataframe.

The Foursquare API was used to obtain all venues in a 2 km radius of each station (figure 4). A list of all stations, with their names, codes and coordinates, can be found

	FECHA	HORA	ACO	AJM	ATI	BJU	CAM	CCA	CHO	COY	...	SAG	SFE	SJA	TAH	TLA	TLI	UAX	UIZ	VIF	XAL
0	2019-01-01	1	-99	-99	-99	-99	-99	-99	-99	-99	...	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
1	2019-01-01	2	-99	-99	-99	-99	-99	-99	-99	-99	...	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
2	2019-01-01	3	-99	-99	-99	-99	-99	-99	-99	-99	...	-99	-99	-99	-99	-99	-99	-99	-99	-99	-99
3	2019-01-01	4	-99	1	15	15	17	15	13	-99	...	15	7	-99	11	14	-99	14	16	18	17
4	2019-01-01	5	-99	1	12	15	16	13	12	-99	...	14	6	-99	7	14	-99	14	16	16	16

Figure 3: Raw data extracted from the NO_2 file. The first column, 'FECHA', is the date. The second column is the hour at which the average concentration of NO_2 was taken, and the remaining columns correspond to the codes of each station (see figure 2). Each -99 value represents missing data.

at the appendix.

	Code	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	AJM	19.2721	-99.207658	Carnitas del Ajusco	19.272523	-99.208687	Mexican Restaurant
1	AJM	19.2721	-99.207658	El Mostachón	19.273790	-99.207150	Taco Place
2	AJM	19.2721	-99.207658	La naranja mecánica	19.273658	-99.207507	Convenience Store
3	AJM	19.2721	-99.207658	Oxxo	19.274657	-99.209401	Convenience Store
4	AJM	19.2721	-99.207658	Sykaryos Paintball	19.272159	-99.209839	Paintball Field

Figure 4: Raw data of venues around the AJM station.

3 Methodology

3.1 Cleaning the data

For the venues dataframe, every duplicate was dropped to prevent overlapping between stations, retaining only the first occurrence. Each different venue category is treated as a categorical variable, and thus, as an extra dimension to our data. Given that our dataset has less than 50 stations, performing statistics on those stations with more than a hundred categories (or dimensions) would yield extremely poor results. Since 2137 venues, spanning more than a hundred venue categories were found, it became necessary to group our venues into 15 general groups to reduce dimensionality. The criteria is as follows. If any venue's category contains a term that exists to the right of the arrows below, then that venue is assigned to the general group to its left. Then, all venues inside each general group are counted and assigned to the stations that the original venues belong to.

- Stores ← ('Store', 'Shop', 'Auto Dealership', 'Bookstore', 'Bakery', 'Market')
- Restaurants & Bars ← ('Restaurant', 'Bar', 'Joint', 'Beer', 'Place', 'Botanero', 'Breakfast Spot', 'Food', 'Brewery', 'Tea', 'Cafeteria', 'Buffet')

- Medical ← ('Dentist', 'Dental', 'Doctor', 'Veterinarian')
- Buildings & Offices ← ('Building', 'Office')
- Hotels & Hostels ← ('Bed & Breakfast', 'Hostel', 'Motel')
- Museums & Galleries ← ('Museum', 'Gallery', 'Exhibit')
- Plazas ← ('Plaza')
- Farms ← ('Farm')
- Parks & Forests ← ('Park', 'Forest')
- schools ← ('University', 'School')
- Gyms & Sports ← ('Gym', 'Sports', 'Studio', 'Martial Arts', 'Paintball', 'Go Kart')
- Stadiums & Theaters ← ('Stadium', 'Concert Hall', 'Theater')
- Bus Stations ← ('Bus Stop', 'Bus Station')
- Sports Fields ← ('Soccer', 'Golf')
- Industrial ← ('Gas Station', 'Distillery')

The resulting dataframe can be seen in figure 5.

	Stores	Restaurants & Bars	Medical	Buildings & Offices	Hotels & Hostels	Museums & Galleries	Plazas	Farms	Parks & Forests	Schools	Gyms & Sports	Stadiums & Theaters	Bus Stations	Sports Fields	Industrial
AJM	7	28	0	1	0	0	0	0	1	0	4	0	0	0	0
ATI	5	22	0	1	0	0	0	0	2	0	4	0	0	2	0
BJU	49	54	0	3	1	2	0	0	4	0	13	8	0	4	0
CCA	9	61	0	0	0	2	1	0	3	2	9	7	2	1	0
CHO	6	29	0	1	0	0	0	0	0	0	3	0	0	0	0

Figure 5: Processed venue data. Every venue was grouped inside a more general group and assigned to its corresponding station. In total, there are 15 general venue categories.

For each pollutant, the maximum concentrations registered were found. Since these concentrations were contained within the winter months (plus march), only the data corresponding to december, january, february, and march of 2019 was kept. The rest of the pollution data was discarded. After this, each dataframe was analyzed to see how many invalid values were in each column, and kept every column that had at least 66% valid data. The rest of the columns were discarded. Then, every remaining column for every pollutant was averaged, producing 'seasonal averages' of each pollutant and for each station. After this, every averaged column was compiled in a new dataframe, and joined with the venues dataframe, so that every station had its corresponding pollutant seasonal averages and its corresponding general groups of venues. Missing data was filled with the mean of its respective column, and noise was added to prevent duplicates,

since clustering algorithms would instantly group together stations with the same values. Next, transforming the pollutants to IMECA units was done according to the Gaceta Oficial del Distrito Federal (the official Mexico City law on how to calculate this). The resulting dataframe, with general venue categories and IMECA scores, can be seen in figure 6

	CO	NO2	O3	SO2	PM10	PM25	Altitude	Stores	Restaurants & Bars	Medical
AJM	26.886450	3.726759	27.180964	4.136824	40.353165	142.950383	2619	7	28	0
ATI	34.443400	5.321372	16.934530	5.287012	51.215565	145.178786	2341	5	22	0
BJU	42.041990	7.877445	19.482670	8.434276	80.090142	159.060493	2250	49	54	0
CCA	34.909223	5.713945	20.618429	3.242201	47.832773	146.795545	2280	9	61	0
CHO	42.693143	4.661582	16.728939	1.436113	54.307063	153.160822	2253	6	29	0

Figure 6: Complete dataset. Note that PM25 oscillates around the 'very bad' IMECA score. The rest of the venue types to the right of 'Medical' are not shown.

Normalization of the dataset was done as follows. For each venue column, the standard MinMax method was used (that is, using a minimum and a maximum value for each column). For the pollutants, a slightly different method was used: only one maximum and only one minimum value for normalization were taken from all the pollutant columns as a whole, and not separately for each column. The main reason for this is that it allows us to preserve the original proportions between pollutants: since they are in the same scale (IMECA units), we can preserve the weights of their contributions to correlation analyses by normalizing them using the same parameters. To represent all IMECA scores in one variable and further reduce dimensionality, every station had its IMECA scores averaged and added to a new column, 'pIndex' ('pollution index'). The resulting column was also normalized using the standard MinMax method (figure 7.)

	pIndex	CO	NO2	O3	SO2	PM10	PM25	Altitude	Stores	Restaurants & Bars	...
AJM	0.063152	0.160576	0.014453	0.162434	0.017040	0.245542	0.892866	0.843750	0.046512	0.250000	...
ATI	0.206851	0.208255	0.024514	0.097785	0.024297	0.314077	0.906926	0.332721	0.000000	0.150000	...
BJU	0.847470	0.256198	0.040641	0.113862	0.044154	0.496257	0.994511	0.165441	1.000000	0.700000	...
CCA	0.214846	0.211194	0.026990	0.121028	0.011395	0.292734	0.917127	0.220588	0.093023	0.766667	...
CHO	0.366518	0.260306	0.020351	0.096488	0.000000	0.333582	0.957288	0.170956	0.023256	0.266667	...

Figure 7: Final dataset. Every value is normalized. Columns to the right of 'Restaurants & Bars' are not shown.

3.2 Clustering

To decide which variables would be used for clustering, a correlation analysis between pIndex and the rest of the variables was performed. Since the strongest correlation of

pIndex was with the Altitude variable (figure 8), Altitude and pIndex were chosen as the two clustering variables.

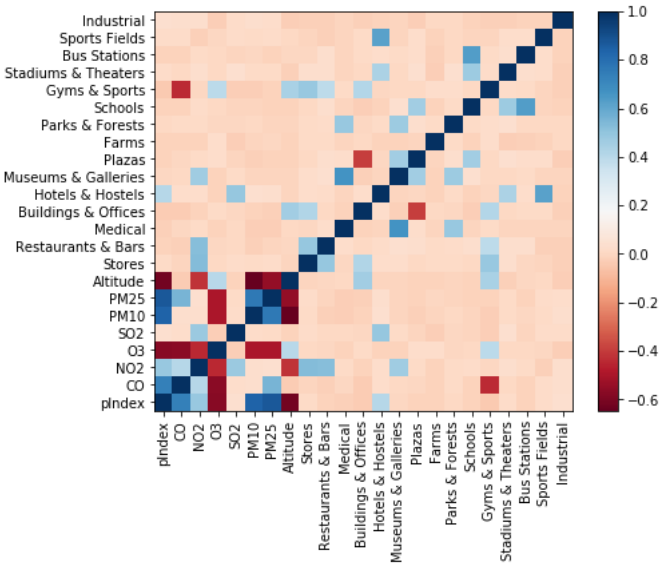

Figure 8: Correlation matrix for GMX. We only clearly see the statistically significant correlations (p-value <= 0.05), since statistically insignificant correlations were masked. Note that pIndex has a moderate, negative correlation with Altitude.

Clustering was done using k-means clustering. Choosing the right amount of clusters was done by testing multiple numbers of clusters. In figure 9, it can be seen that the elbow that corresponds to the maximum silhouetter score is at 5 clusters; however, using 5 clusters results in some clusters having only two stations. Since correlations with two samples always yield either +1 or -1 correlations, from which no useful information can be inferred, a minimum of three stations per cluster is needed.

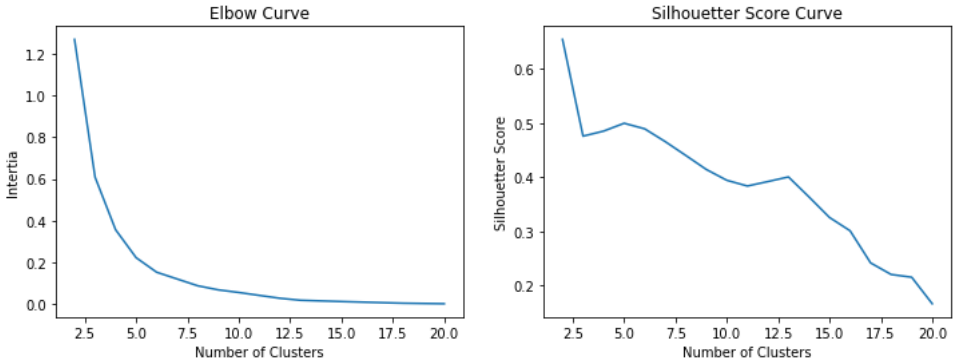

Figure 9: Elbow curve (left) and silhouetter scores (right) for multiple amounts of clusters.

4 Results & Discussion

By using 4 clusters instead of 5, the clusters in figure 10 are obtained.

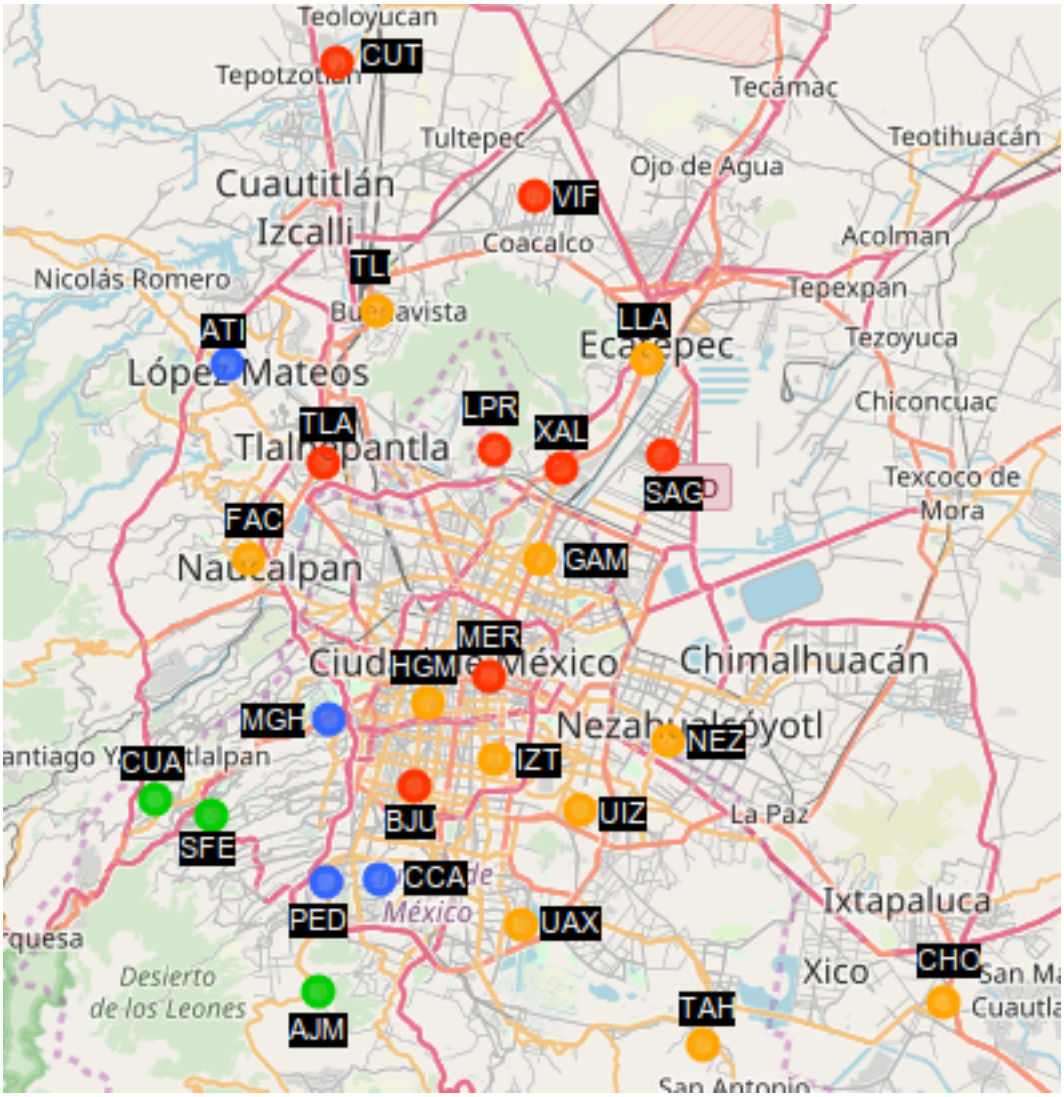

Figure 10: Labeled clusters in GMX. The green dots (cluster 2) are the stations with the smallest pIndex and highest Altitude, followed by the blue (cluster 3), orange (cluster 1) and red dots (cluster 0), which are the most polluted stations.

4.1 Cluster 0

	pIndex	CO	NO2	O3	SO2	PM10	PM25	Altitude	Stores	Restaurants & Bars	Medical
BJU	0.847470	0.256198	0.040641	0.113862	0.044154	0.496257	0.994511	0.165441	1.000000	0.700000	0.0
CUT	0.706593	0.351886	0.021723	0.106789	0.025036	0.379847	0.979027	0.189338	0.069767	0.083333	0.0
LPR	0.741839	0.440067	0.028166	0.109363	0.011614	0.337304	0.958139	0.261029	0.209302	0.016667	0.0
MER	0.622147	0.290442	0.042524	0.107401	0.023314	0.385396	0.966489	0.156250	0.581395	0.666667	0.0
SAG	0.645533	0.295999	0.028343	0.093548	0.012488	0.429552	0.969134	0.148897	0.139535	0.333333	0.0
TLA	0.616621	0.317312	0.038944	0.090505	0.036439	0.369252	0.959923	0.277574	0.418605	1.000000	0.0
VIF	0.607153	0.227209	0.022628	0.089884	0.034898	0.458350	0.973940	0.150735	0.069767	0.233333	0.0
XAL	1.000000	0.370503	0.040105	0.081705	0.021105	0.520247	1.000000	0.000000	0.162791	0.083333	0.0

Figure 11: Every station in cluster 0. The columns to the right of Medical are not shown.

These stations (figure 11, red dots in figure 10) report the highest pollution. Understandably, the industrial category has a very strong, positive influence on pollution here (figure 12), as the only station with a factory nearby (XAL) has the highest pIndex out of all the stations in the dataset; however, we have to consider that this is one of the caveats of using extremely small sized sets of samples: every individual sample will have an enormous effect on whatever statistics we derive. Before we're content with our confirmation that industries pollute, we should see every correlation in our cluster.

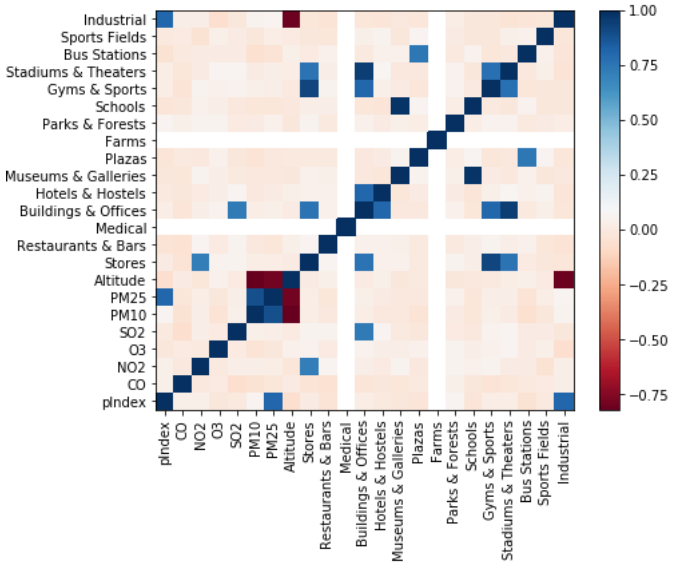

Figure 12: Correlation matrix for cluster 0.

The Industrial variable has a strong, negative correlation to altitude, and altitude has strong, negative correlations with PM25 and PM10; PM25 has a strong, positive correlation to pIndex. We can conclude that the Industrial variable is indeed strongly

correlated to pIndex, but not by causation, but by association, since a low altitude, in this particular case, is associated both with the presence of a soap factory and PM25, which is strongly correlated to pIndex. If we had more than one instance of the Industrial variable in this cluster, we could analyze this further. Other than that, there are no other venue types that are directly correlated to pIndex. Stores are correlated to NO_2 and Buildings & Offices could be correlated to SO_2, but its regression plot (figure 13 (d)) tell us that it probably isn't.

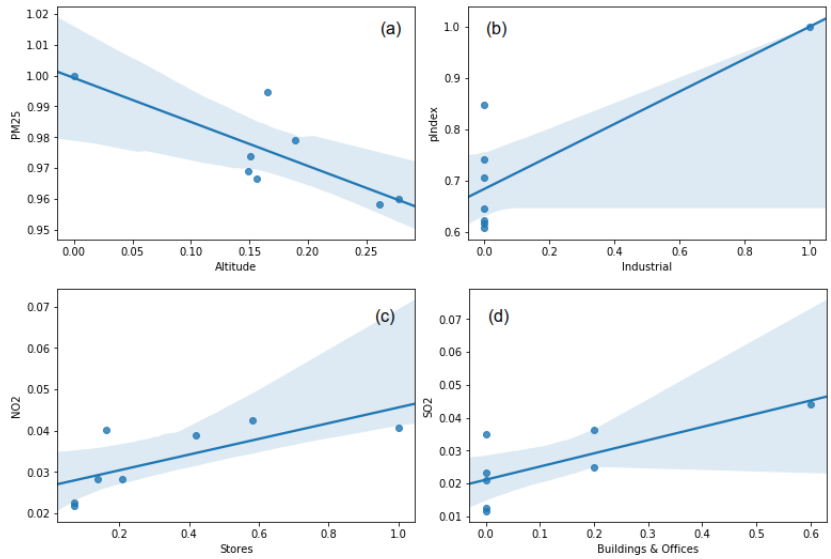

Figure 13: Regression plots of interest for cluster 0.

4.2 Cluster 1

This cluster (figure 14, orange dots in figure 10) represents the moderately polluted areas. This is an understatement, however: let's not lose sight of PM25, which is very high across every cluster, oscillating about the "very bad" IMECA level.

Here, we have bus stations only near the GAM station, and this seems to be negatively correlated (figure 15) to CO; however, this is the same case as in the Industrial variable on cluster 0. We will not trust that correlation. Fortunately, there are Restaurants & Bars and Gyms & Sports venues near every station in this cluster, and those two variables are correlated to different pollutants (even if not directly to pIndex).

Every regression plot here (figure 16) behaves kind of like a wave. However, in plots (d) and (e), it looks like some specific points are tilting the correlations to make them seem stronger than they really are. For example, in plot (e), if we took the first point away, the correlation between PM25 and Gyms & Sports would look extremely weak, or even nonexistent. I would argue that the Gyms & Sports venues are only correlated to CO, SO_2 and O_3, due to the fact that air quality is directly related to the amount of exercise a person can do, and thus, areas with lower pollution are more suited for

	pIndex	CO	NO2	O3	SO2	PM10	PM25	Altitude	Stores	Restaurants & Bars	Medical
CHO	0.366518	0.260306	0.020351	0.096488	0.000000	0.333582	0.957288	0.170956	0.023256	0.266667	0.0
FAC	0.430544	0.299086	0.032952	0.119550	0.026691	0.269352	0.957341	0.255515	0.186047	0.716667	0.0
GAM	0.393242	0.171998	0.030883	0.126453	0.002680	0.378658	0.972767	0.123162	0.534884	0.550000	0.0
HGM	0.439562	0.254764	0.037537	0.113954	0.021208	0.315251	0.967460	0.136029	0.627907	0.783333	0.0
IZT	0.486104	0.289204	0.036225	0.108617	0.015849	0.325170	0.961974	0.143382	0.441860	0.950000	0.0
LLA	0.469941	0.285730	0.030699	0.098515	0.021647	0.333786	0.957334	0.128676	0.488372	0.133333	0.0
NEZ	0.568160	0.292895	0.028917	0.119268	0.018692	0.345143	0.979488	0.137868	0.209302	0.383333	0.0
TAH	0.471490	0.212784	0.015414	0.138732	0.002397	0.388521	0.970757	0.251838	0.000000	0.000000	0.0
TLI	0.391678	0.293454	0.032427	0.118381	0.048196	0.261401	0.928677	0.281250	0.162791	0.133333	0.0
UAX	0.423437	0.236550	0.022073	0.110762	0.006811	0.365995	0.958677	0.158088	0.558140	0.616667	0.0
UIZ	0.502548	0.271620	0.031608	0.114503	0.011488	0.349561	0.967753	0.112132	0.162791	0.733333	0.0

Figure 14: Every station in cluster 1. The columns to the right of Medical are not shown.

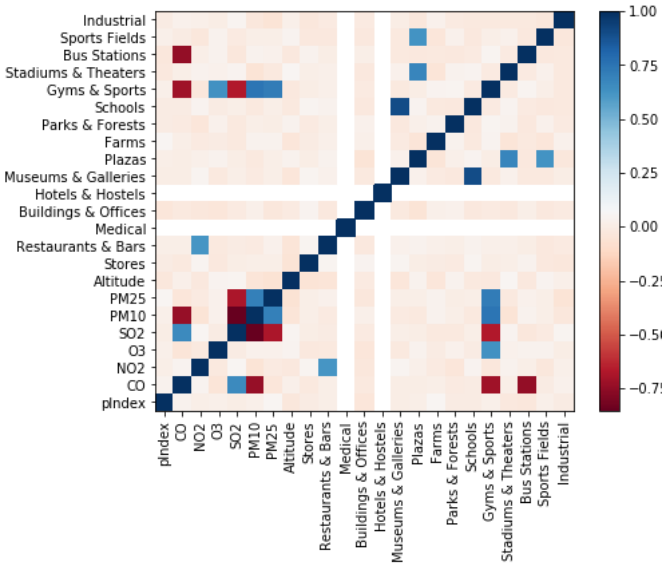

Figure 15: Correlation matrix for cluster 1.

exercise. As for the wave-like behavior, more data is needed to explain it.

Even though the relationship between Restaurants & Bars and NO_2 might not be linear (figure 16 (f)), it's clearly positive. One source of NO_2 is the combustion of fossil fuels, such as coal, gas and oil. Having many places to eat, in a place such as GMX, guarantees the existence of many taco places and burger joints. This correlation, then, could be more of a causation.

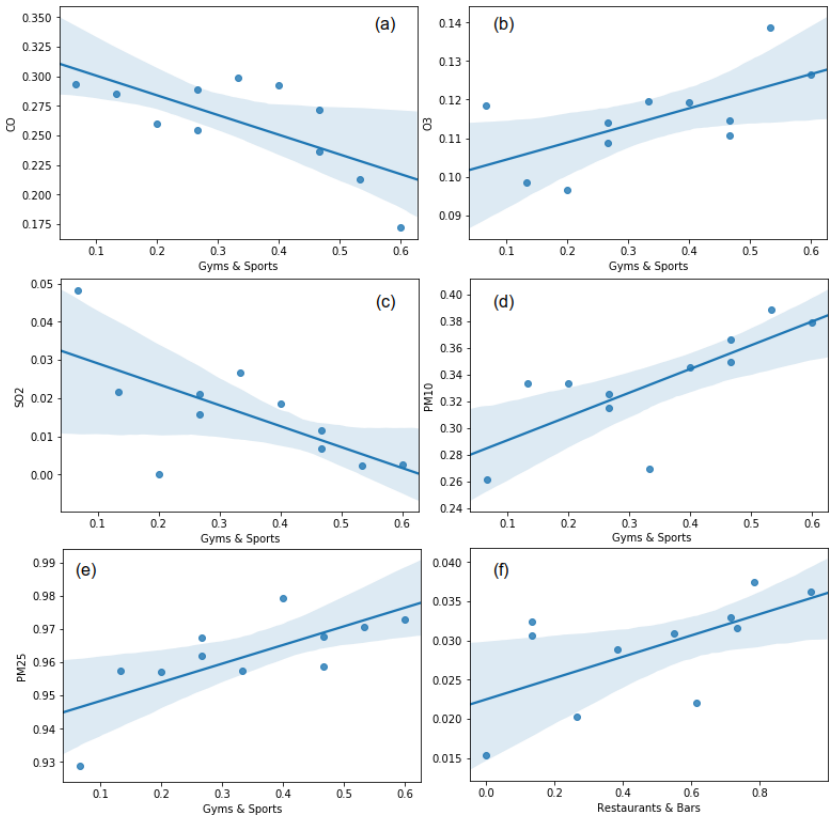

Figure 16: Regression plots of interest for cluster 1. Most regression plots look like sine waves.

4.3 Cluster 2

This cluster (figure 17, green dots in figure 10) includes the stations with the highest altitude and the least pollution. There are only three stations, so we need to be extra wary of what our statistics tell us.

	pIndex	CO	NO2	O3	SO2	PM10	PM25	Altitude	Stores	Restaurants & Bars	Medical
AJM	0.063152	0.160576	0.014453	0.162434	0.017040	0.245542	0.892866	0.843750	0.046512	0.250000	0.0
CUA	0.303390	0.294866	0.023970	0.109620	0.012078	0.224163	0.966879	1.000000	0.465116	0.433333	0.0
SFE	0.000000	0.153873	0.025254	0.122352	0.007593	0.246152	0.901235	0.806985	0.372093	0.600000	0.0

Figure 17: Every station in cluster 2. The columns to the right of Medical are not shown.

The Parks & Forests variable seems to have a negative correlation with CO (figure 18). Indeed, it is known that plants use CO for many processes [5, 6], which could actually mean that this correlation is a causation. The correlations are consistent, too: Parks & Forests is negatively correlated to CO and positively correlated to PM10, and

PM10 is negatively correlated to CO. The link between these three variables could be the Altitude variable. The higher you go, the less protection you have from sunlight. Sunlight catalyzes many chemical reactions, some of which could be forming PM10 from CO. However, we do not have enough data to confirm this.

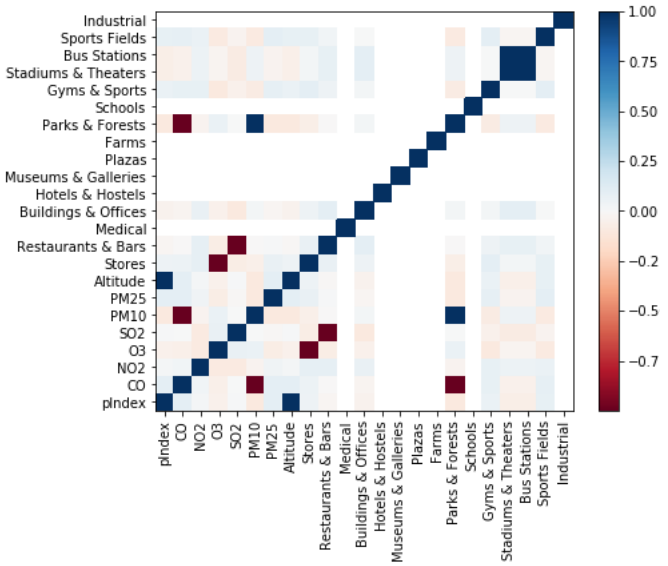

Figure 18: Correlation matrix for cluster 2.

4.4 Cluster 3

Cluster 3 (figure 19, blue dots in figure 10) is yet another small-sized set. Stations here have both low Altitud and low very pIndex.

	pIndex	CO	NO2	O3	SO2	PM10	PM25	Altitude	Stores	Restaurants & Bars	Medical
ATI	0.206851	0.208255	0.024514	0.097785	0.024297	0.314077	0.906926	0.332721	0.000000	0.150000	0.0
CCA	0.214846	0.211194	0.026990	0.121028	0.011395	0.292734	0.917127	0.220588	0.093023	0.766667	0.0
MGH	0.220160	0.266887	0.034691	0.099863	0.020461	0.245423	0.916211	0.378676	0.255814	0.866667	1.0
PED	0.061857	0.173215	0.023902	0.152341	0.010437	0.235708	0.896560	0.305147	0.604651	0.433333	0.0

Figure 19: Every station in cluster 3. The columns to the right of Medical are not shown.

The variables Farms and Medical occur only one station at a time, which is why we should ignore their correlations. The only variables that we can trust are really correlated to NO_2 are Parks & Forests and Museums & Galleries.

The mechanism for which Museums & Galleries would be (weakly) correlated to a NO_2 increase could be attributed to them being places that many people visit all year long. Some of the most iconic museums, galleries and theaters are congregated here,

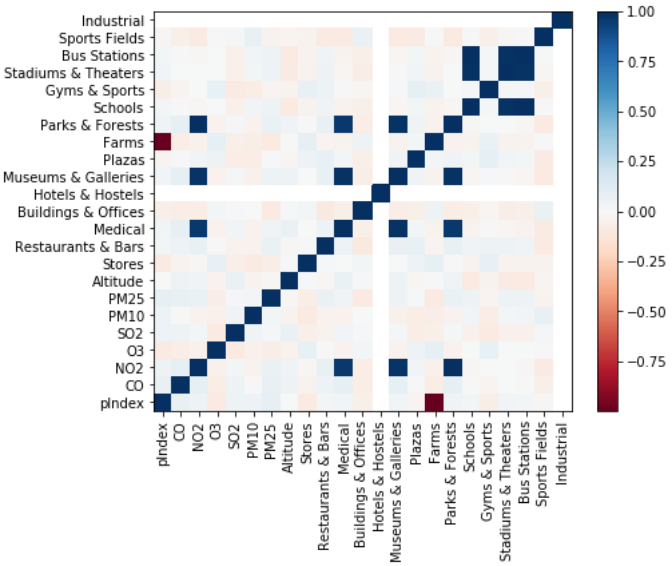

Figure 20: Correlation matrix for cluster 3.

including the National Auditorium. The same can be said about the Parks & Forests variable. The MGH station is located near Chapultepec, a forest which is a major cultural asset in Mexico City. More can be said about NO_2 and forests. A sweden paper [7] suggests that urban forests in heavy-traffic areas have a very small effect in reducing NO_2 and no effect on O_3, while research conducted on the USA and China [8, 9] suggest that O_3 but not NO_2 is reduced by urban trees. Another study, conducted in the Netherlands [10] found that soils in forests can be a net source of reactive nitrogen oxide (NO). NO can be oxidized to NO_2 by reacting with O_3 and other chemical species. From the varying results of research in different countries, we see that urban trees have different effects depending on their specific contexts and species. It could be the case that the Chapultepec forest is actually one of those NO producing forests. Specific studies should be conducted before reaching this conclusion, but we are well justified in suspecting this.

5 Conclusions

While some effects cannot be seen in the complete GMX dataset, they can be seen once we partition our data into smaller datasets. Some of these effects include:

- In the most polluted, lowest altitude regions (cluster 0), the presence of a soap factory nearby the XAL station was strongly correlated with pIndex. We cannot directly conclude that this factory is the culprit, since we do not have many factories in our data, and thus, this station is an outlier inside its own cluster. However, the lack of factories in the data hints that the Foursquare API is should not be used alone for this task.

- In regions with moderate pollution and low altitude (cluster 1), having many places to eat is correlated to higher NO_2 concentrations, and having places where you can exercise is correlated to lower SO_2 and CO concentrations, and higher O_3 concentrations. Thus, these are places where NO_2 emissions are prone to be increased if more places to eat appear.

- About clusters 2 and 3, I suspect that urban forests play different roles. In cluster 2, we see a negative correlation between Parks & Forests and CO. In cluster 3, a positive correlation between NO_2 and Parks & Forests can be seen. While we cannot be sure of the inner workings behind these observations, these results hint at what differing results from different studies imply: that urban forests behave one way here, and a different way over there. Specific research should have to be done to address this.

Also, because of the existence of wind and its undeniable role in pollution transport [11], there might be a correlation between venue and station locations. By looking at our clustering map, we see that the most polluted areas are to the northeast of GMX, and the least polluted to the southwest. This model doesn't consider winds, but our clustering suggests that it should, which is an important conclusion by itself.

6 Final remarks

Through our clustering analysis, we learned about some correlations that were invisible when looking at the whole dataset. This is a double-edged sword, because reducing the size of the datasets will result in more drastic effects being observed, but less generalization power. Ideally, I would have liked to work with a much bigger dataset. Even a hundred stations would have been enough, but after cleaning the data, we ended up discarding about 12 stations, and only 26 remained. Even so, because we clustered our stations using educated criteria, we can trust that any phenomenon observed (such as the positive correlation between forests and NO_2) is well justified in arising suspicion. While our datasets were too small to conduct comfortable statistics, those statistics are still statistics, and can be valuable when interpreted under the right context.

When I get the chance, I would like to repeat this experiment incorporating distances from every venue to every station, since pollution is known to travel by wind. It isn't static, as this model wordlessly implies, and the pollution from some venues could be impacting on venues far away. This would also solve many NaN problems, since every station would have at least a tiny amount associated to each type of venue, which would help the correlation analysis immensely.

Finally, I thank you, the reader, for making it this far, and hope you had as much fun reading this as I had writing it.

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*

[5] Wang, M., & Liao, W. (2016). Carbon Monoxide as a Signaling Molecule in Plants. Frontiers in plant science, 7, 572. doi: 10.3389/fpls.2016.00572

[6] Bidwell, R. G. S. and Fraser, D. E., Carbon monoxide uptake and metabolism by leaves. Canadian Journal of Botany. Vol. 50, 7, 1435-1439. 1972.

[7] Grundström, Maria & Pleijel, Håkan. (2014). Limited effect of urban tree vegetation on NO2 and O3 concentrations near a traffic route. Environmental Pollution. 189. 73–76. doi: 10.1016/j.envpol.2014.02.026.

[8] Yli-Pelkonen, Vesa & Viippola, Viljami & Rantalainen, Anna-Lea & Zheng, Junqiang & Setälä, Heikki. (2018). The impact of urban trees on concentrations of PAHs and other gaseous air pollutants in Yanji, northeast China. Atmospheric Environment. 192. 151-159. doi: 10.1016/j.atmosenv.2018.08.061.

[9] Vesa Yli-Pelkonen & Anna A Scott & Viljami Viippola & Heikki. Trees in urban parks and forests reduce O3, but not NO2 concentrations in Baltimore, MD, USA. 2017.

[10] Duyzer, J. H. & Dorsey, J. R. & Gallagher, M. W. & Pilegaard, K. & Walton, S. Oxidized nitrogen and ozone interaction with forests. II: Multi-layer process-oriented modelling results and a sensitivity study for Douglas fir. Quarterly Journal of the Royal Meteorological Society, 130, 600, 1957-1971. doi: 10.1256/qj.03.125.

[11] Rafael Silva-Quiroz, Ana Leonor Rivera, Paulina Ordoñez, Carlos Gay-Garcia, Alejandro Frank. Atmospheric blockages as trigger of environmental contingencies in Mexico City. Heliyon, Volume 5, Issue 7, 2019, ISSN 2405-8440.

A List of stations

	Borough	State	Latitude	Longitude	Altitude
AJM	Tlalpan	CDMX	19.272100	-99.207658	2619
ATI	Atizapán de Zaragoza	Estado de México	19.576963	-99.254133	2341
BJU	Benito Juárez	CDMX	19.371612	-99.158969	2250
CCA	Coyoacán	CDMX	19.326200	-99.176100	2280
CHO	Chalco	Estado de México	19.266948	-98.886088	2253
CUA	Cuajimalpa de Morelos	CDMX	19.365313	-99.291705	2704
CUT	Tepotzotlán	Estado de México	19.722186	-99.198602	2263
FAC	Naucalpan de Juárez	Estado de México	19.482473	-99.243524	2299
GAM	Gustavo A. Madero	CDMX	19.482700	-99.094517	2227
HGM	Cuauhtémoc	CDMX	19.411617	-99.152207	2234
IZT	Iztacalco	CDMX	19.384413	-99.117641	2238
LLA	Ecatepec de Morelos	Estado de México	19.578792	-99.039644	2230
LPR	Tlalnepantla de Baz	Estado de México	19.534727	-99.117720	2302
MER	Venustiano Carranza	CDMX	19.424610	-99.119594	2245
MGH	Miguel Hidalgo	CDMX	19.404050	-99.202603	2366
NEZ	Nezahualcóyotl	Estado de México	19.393734	-99.028212	2235
PED	Álvaro Obregón	CDMX	19.325146	-99.204136	2326
SAG	Ecatepec de Morelos	Estado de México	19.532968	-99.030324	2241
SFE	Cuajimalpa de Morelos	CDMX	19.357357	-99.262865	2599
TAH	Xochimilco	CDMX	19.246459	-99.010564	2297
TLA	Tlalnepantla de Baz	Estado de México	19.529077	-99.204597	2311
TLI	Tultitlán	Estado de México	19.602542	-99.177173	2313
UAX	Coyoacán	CDMX	19.304441	-99.103629	2246
UIZ	Iztapalapa	CDMX	19.360794	-99.073880	2221
VIF	Coacalco de Berriozábal	Estado de México	19.658223	-99.096590	2242
XAL	Ecatepec de Morelos	Estado de México	19.525995	-99.082400	2160

Figure 21: Complete list of surviving stations after data cleanup.