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 a solid candidate to study pollution. It contains heavily urbanized areas, rural areas, mountains and basins. Thus, geography, varying wind speeds, sunlight, and the presence and interaction of multiple gasses, which come from both natural and artificial sources, make analyzing the mechanisms behind pollution a daunting task. A deeper understanding of pollution needs to be achieved before implementing new - and possibly very costly - ways to reduce it.

2 Data

The air quality accross all of GMX is being measured by the government in 45 fixed monitoring stations. There are six pollutants of interest and for which data is available [4]: carbon monoxide (CO), sulfur dioxide (SO_2), ozone (O_3), nitrogen dioxide (NO_2), particulate matter less than 10 microns in diameter (PM_{10}) and particulate matter less than 2.5 microns in diameter ($PM_{2.5}$). The data consists of hourly averages of the measured pollutants, for each day of year, for all 45 stations.

The Foursquare API will be used to obtain data about the surroundings of each station, and statistics will be used to determine correlations between pollution and the amounts and types of venues using the entirety of the venue and air quality data of GMX. Then, depending on these correlations, the stations will be clustered using k-means clustering, and correlations between pollutants and venues will be determined again, separately, for each cluster.

It is expected, given the environmental diversity of GMX, that each cluster will present a different set of correlations. Once these correlations are established, we will discuss them under the lens of physics and chemistry.

Note that the goal of this work is not to propose new ways of countering pollution, but to gain some more insight, which hopefully, would be useful in further, more specialized studies.

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*