Universidad Autonoma de Nuevo Leon
Facultad de Ingenieria Mecanica y Electrica
Posgrado de Ingenieria de Sistemas

RESEARCH PROJECT PROPOSAL

Name of Candidate: Jose Alberto Benavides Vazquez

Date of Admission to this programme: September 9, 2022

Full time mode

1. Project Title

Impact of the air pollutants in the population health in the Monterrey Metropolitan Area.

2. Introduction

This study focuses on the impact that the air quality have in the population welfare primarily in the Monterrey Metropolitan Area between 2018 and 2019 obtained from the Comprehensive Environmental Monitoring System (abbreviated SIMA in Spanish from Sistema Integral de Monitoreo Ambiental) [**?**]. Also, it have been incorporated a database containing the hospital discharge records for Mexico over the same span of time from the open repository of the Department of Health of Mexico [**?**].

As the atmosphere is result of the combination of 78% of N_2, 21% of O_2 and 1% of combined helium, water vapor, oxides, methane, noble gases, sulfides, ozone, ammonia, formaldehyde, and hydroxyl radical, can be defined that the air is *polluted* when these concentrations are altered by human activities resulting in harmful environments to biological ecosystems [**?**].

The human activities that cause pollution have been increasing over the years. Air quality became an urban problem since 1930 when a fog in Belgium was associated with the death of 63 people and then in London during 1952 [**?**] when another fog caused $4,000$ deaths [**?**] being pneumonia the principal illness diagnosed in both cases. Its negative impact on health and general welfare concerned authorities that design regulations in order to minimize those undesirable effects. In 1981 it became an international problem when the CO concentrations were seen from space for the first time [**?**]. Such images show that the pollution was not only generated from expected sources, but from other unforeseen activities, and that the pollution generated in one place affected distant populations over time.

Air pollution can be caused by events that impact the atmosphere, like forest fires, increased volcanic activity, animal and vegetable decomposition, climate change [**?**], etc. The main difference from natural and human pollutants is that the natural ones tend to occur far from large human populations and are less often than human sources of pollution, which are generally located in or near large human populations, and their emissions are more frequent. The main human sources of air pollution are associated with transportation, fuel combustion, and industrial processes [**?**].

The International Classification of Diseases (ICD)[1] is a code associated to diseases by their trends and statistical values, established by the **?** [**?**]. It consists, generally, in a letter and three numbers between 0 and 9. For example, the code for `pneumonia due to staphylococcus` is `J152`. This system comprehends from the values `A00.0` to `Z99.9`, leaving the `U` letter for yet unknown sources. This classification was proposed by Farr at the end of the XIX century [**?**],

[1] The name of the diseases corresponding to the ICD code is obtained from https://icd.who.int/browse10/2019/en

and its purpose is to understand the causes of morbidity and mortality in order to improve the quality of life of the human population [?].

3. Hypothesis

The contamination levels reported can explain the statistically significant variations and trends of georeferenced data from health reports. Also, it is possible to automatize these relationships in an interactive and free web service.

4. Project objectives

- Establish a clear and scientifically sustained relationship between air pollution and the human diseases that them can cause.
- Study the impact of the air pollution in the Metropolitan Monterrey Area during 2018 and 2019 in the health population.
- Choose or propose a methodology to calculate the causation between air pollution and associated human diseases.
- Identify the areas where they are more air pollution and establish a relationship between the people that live nearby those areas.
- Propose a rank to determine which air pollutants have the most direct effect in the human health.
- Propose pollution emission limits, scientifically justified, that can be used by organizations interested in this matter.
- Generate interactive and free tools that let people and specialist know the levels, relationships, risks, etc. related to the air pollutants studied.

5. Research questions

- Is it possible to establish statistically significant relationships between air pollution and human diseases?
- What are the best methods to detect the causation between air pollution and human diseases?
- How to eliminate the climate and seasonal components from the relationship between air pollution and human diseases?
- Does the environmental and geographical factors can be used to predict the amount of air pollution in an area and time?
- Which air pollutants cause more human diseases?
- Is it possible to establish short (days), medium (months) and large (years) interactions between air pollution and human diseases?

6. Background and related literature

The first correspondence to discuss relationship between air quality and human health was registered during December 1930 in Belgium, when a three-day fog s was declared to be the cause of death of 63 people — during the fog, disease and death were observed to increase, whereas after its dissipation, the figures normalized again [?]. Similarly, in December 1952, a four-day fog in London was associated with four thousand deaths [?]. The majority of the fog-related deaths were related to respiratory diseases.

In ? ? [?] used a method to quantify the severity of respiratory illness based on their condition (negative numbers indicated they were recovered and positive numbers that they were getting worse). Then, he plotted a time series superposed with concentrations of two pollutants — SO_2 and smoke — and two weather variables — temperature and humidity; he discovered that the pollutants bore a similarity to the severity whereas the climate ones did not.

During the 1970s, short-term exposure to different pollutants was studied [? ? ? ?] with experiments that implied some ethical and legal considerations, focusing on physiological analysis of the subjects. A compilation was published in by ? [?] who documented that the pollutants, in regulated concentrations, had a negative impact on health, primarily on respiratory diseases.

A set of methodologies are used to measure the relationships that are the focus of this study. They are **multiple regression analysis** [? ? ? ?] with variations in the distributions such as logistic regressions [?], **multivariate analysis** [? ?], **auto regressive models** [?], **causality models** [?] and **case-crossover approximations** [? ?].

7. Methodology

The study will focus in the **Monterrey Metropolitan Area (MMA)** between 2018 and 2019. The air quality samples in the are taken each hour from 13 sensors over the MMA that measure concentrations of **CO**, **NO**, **NO**$_2$, **O**$_3$, **SO**$_2$, **PM**$_{10}$, **PM**$_{2.5}$, and **atmospheric pressure**, **rainfall**, **relative humidity**, **solar radiation**, **temperature**, **wind velocity and direction**. This data is provided by the *Sistema Integral de Monitoreo Ambiental de Nuevo León* (SIMA) [?]. The Mexican diseases data was obtained from the Mexican *Department of Health* [?] and contains information from all states and municipalities in Mexico such as **date of admission**, **egress date**, **age at the admission**, **gender**, **weight**, **height**, **ICD code upon arrived**, **ICD code upon diagnosis**, and **reason of egress**.

The air quality data needs to be interpolated because it contains imputed records. Different temporal interpolation techniques are used and compared [?]. Also, an spatio-temporal interpolation is performed to obtain the missing data values [?]. Both data sets are processed and converted to georeferenced time series [?] that are stationary [?] in order to establish their relationship by cross correlation [?], multiple regression analysis [?], vector autoregressive approaches, causality models [?] and geographic interactions [?]. The results are ranked by metrics like R^2, the Akaike (AIC), and the Bayesian (BIC) information criteria [?]. Finally, it will be produced a web application that allow general and specialized population to interact with the data and obtain forecasts, interactions and visualizations of the models described.

8. Expected results

Univariate and multivariate analysis are performed on the time series from both the data sources of air quality and human diseases variables. Descriptive statistical results are expected to inform levels of pollutants and its levels according to local legislation (shown in Figure 1) or timespans where more cases are reported (found in Figure 2). Causal analysis are based in causal diagrams (like the one on Figure 3). The causal tests have confidence intervals where $\alpha = 0.05$. For example, Figure 4 shows the statistically significant results of a Granger test on x causing y variables in green. When there is a causal relationship between variables, mathematical models are able to find the relationships between the variables. Here, a vector autoregressive model is used to calculate the interactions between variables and its temporal lags. This results in equations where a variable can be explained from the most significant interactions and its coefficients. As an example, the `J` diseases depends on the interactions

$$Y_J = -1.96a_{\text{J:1}} - 3.81a_{\text{J:6}} - 0.89a_{\text{PM10:14}} - 6.71a_{\text{J:7}} - 1.31a_{\text{PM10:6}} \dots$$

where a are the values of the variables and lags written as subindex in the form `variable:lag`.

9. Significance

The relationship of air pollutants and human diseases have been widely studied. From those studies it is evident that the forecasting that can be produced it is relevant to the scientific community and other decision making agents. Nevertheless, the causation interactions of those factors are primarily studied in focus groups experiments where the ethical concerns limits the exploration. Mathematical causality models can bring a better understanding of the network of interactions between pollutants, diseases and confounding factors like weather conditions such as temperature, atmospheric pressure and so on, supporting the possibility of simulations that extends the ethical concerns limits discussed.

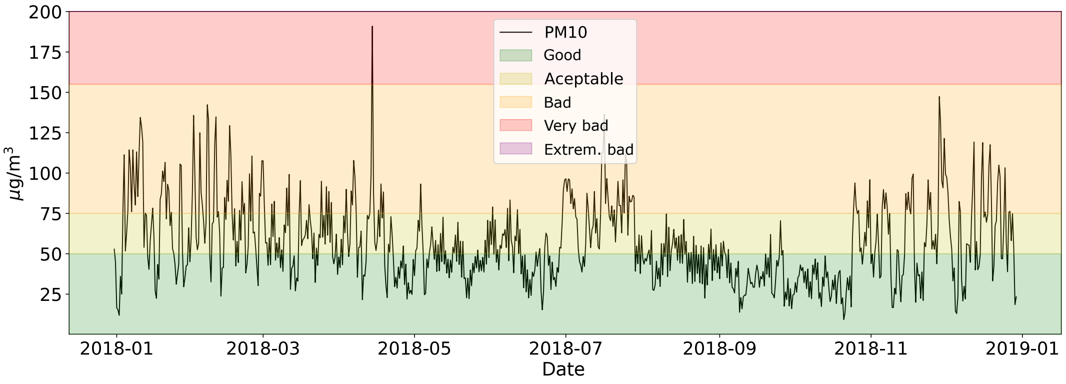

Figure 1: PM_{10} 12-hour mean concentration for the MMA in 2018 displayed over the categorical ranges.

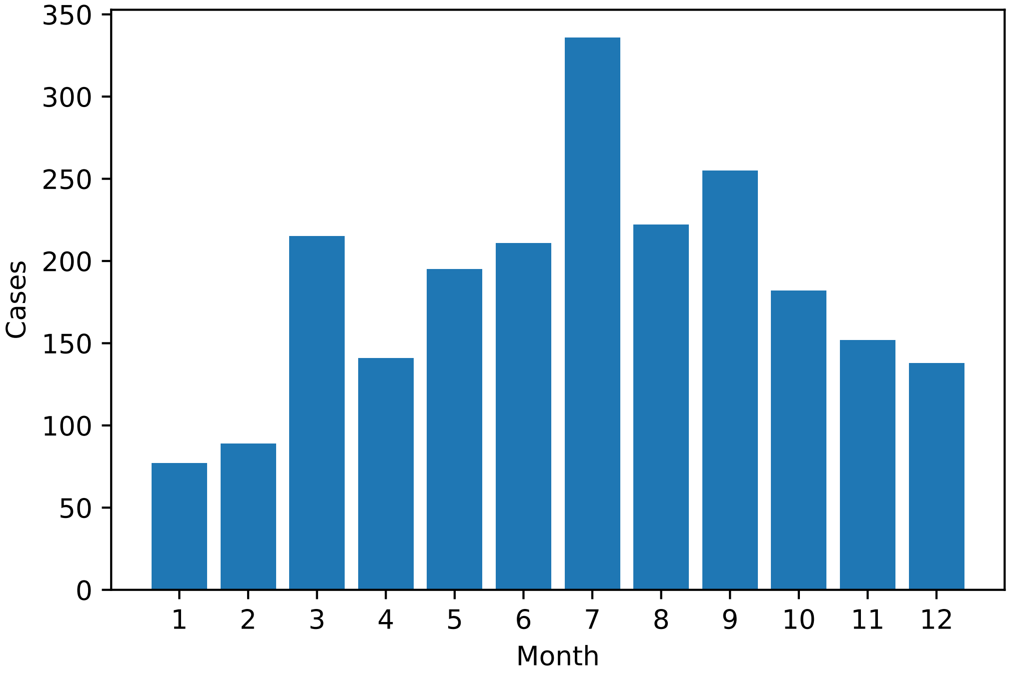

Figure 2: Bar diagram of the preprocessed records of patients from the MMA in 2017.

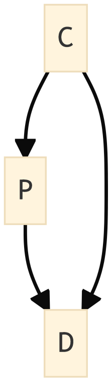

Figure 3: Example of a causal diagram where pollutant P cause disease D, and both are caused by a confounding factor C.

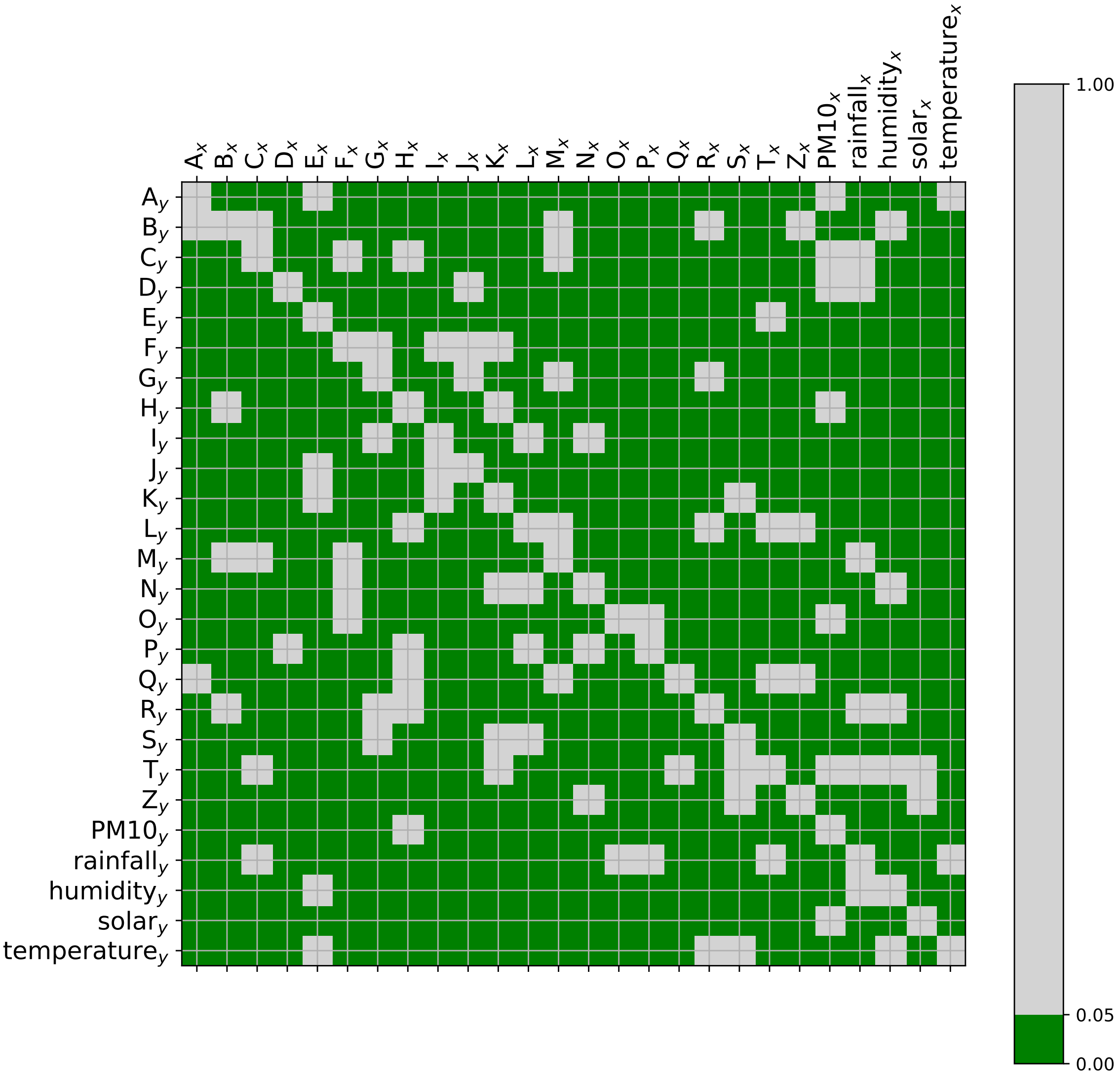

Figure 4: Granger causality test for x (horizontal) variables causing y variables (vertical).

10. Project schedule

The Gantt diagram of this project can be accessed in https://aulaplus.notion.site/c5e2f4f3b8004123ac356811e598a63e?v=34166671ce074e6499778146b40b3462.