Machine Learning

Lucas Machado Moschen

School of Applied Mathematics,
Fundação Getulio Vargas

May 26, 2021

Abstract

Air pollution is one of the contemporary most important topics.

Todo list

1.

1 Introduction

Rio de Janeiro city (Brazil) is one of the most beautiful cities in the world, according to the travel website Conde Nast Traveler [2]. This thought is shared among tourists and residents.

1. Environmental pollution is a consequence of the development of cities and they can cause damage on human health.
2. Fossil fuels, agriculture, industries, natural disasters are the main causes of air pollution.
3. The program MonitorAR exists since 2011.
4. Find data about deaths in Rio de Janeiro.
5. What is your basic approach? : This study aims to build models for hourly/daily air quality forecasting for the city of Rio de Janeiro, using the algorithms XXX and YYY.
6. The text is organized as follows...

2 Problem definition

1. What is the problem?
2. What are the inputs and outputs mathematically.

3 Background of air pollution

Air pollution is a mixture of particles and gases, often not visible to human eyes. The visible forms are widely known, such as smoke, soot and mold. The quality air index, calculated based on the levels of the gases, is available in a diary report, as presented in image XXX.

3.1 Polluting gases

The atmosphere of the Earth is a dynamic and complex system of natural gases, which are necessary to life [3]. The planet has a defense mechanism that

absorbs part of these fases, what forms a cycle. However, high levels of gases concentration can cause several effects in the living beings. The polluting gases include:

- **Óxidos de Carbono:** O monóxido de carbono (CO) é oriundo da combustão incompleta e não apresenta cheiro ou cor. Já o dióxido de carbono é um gás que contribui para o efeito estufa e, em excesso na atmosfera, devido à queima de combustíveis fósseis, pode causas sérios danos.
- **Óxidos de Nitrogênio:** Também emitidos por veículos e tem uma aparência marrom. O dióxido de nitrogênio é um dos gases mais perigosos para a poluição do ar, e sua toxidade é facilmente identificável.
- **Óxidos de Enxofre:** Causa primária da chuva ácida, muito comum na Europa. É natural após erupções vulcânicas. É uma forte causa de problemas respiratórios.
- **Ozônio:** O gás ozônio (O_3) contém três átomos de oxigênio. Até pequenas concentrações desse gás são consideradas tóxicas e explosivas. Ele ocorre naturalmente na atmosfera, porém em pequenas quantidades, quando absorve radiação ultravioleta. Em condições especiais, óxidos de nitrogênio e hidrocarbonos podem produzir ozônio em concentração alta o suficiente para causar irritação nos olhos e na mucosa.

4 Methodology

1. Criteria to evaluate the methods.
2. What hypotheses am I testing?
3. Experimental methodology.
4. Dependent and independent variables.
5. Train and test data
6. Comarisons to competing methods

5 Exploratory data analysis

5.1 Data description

The dataset used in this study was extracted from the project MonitorAr [1]. The table contains hourly data observations, separated by pollutant, weather condition, and monitoring stations' characteristics from the city of Rio de Janeiro. Table 1 informs the most important variables used, and Table 2 indicates the measured pollutants. The events were collected between January 1, 2011, and March 31, 2021. A total of 661,662 records were used.

1. Reportar valores nulos da chuva e índice de missing values

1. Gráficos dos gases e interpretação de alguns deles. Analisar curtose e assimetria.
2. Mensurações temporais de alguns gases. Selecionar alguns poucos
3. Testes de estacionariedade nas séries utilizadas.
4. Mais alguns gráficos de visualização.

5.2 Data preprocessing

The data preprocessing is an important step before the usage of machine learning algorithms, in order to report robust and neat results.

1. Imputation of missing data
2. Handling outliers
3. normalization and standardization.
4. feature engineering

5.2.1 Missing data imputation

In this dataset, there is two types of missing data: (1) monitoring stations do not measure all pollutants by construction. For instance, it is not measured NOx in Centro and Copacabana; and (2) monitoring stations did not measure in a period for some reason. We have to deal with them in two different ways.

	Name	Type	Description
Meterological conditions	Chuva	float	Rainfall (mm)
	Pres	float	Atmospheric Pressure (mbar)
	RS	float	Solar radiation (w/m2)
	Temp	float	Temperature (°C)
	UR	float	Relative humidity (%)
	Dir_Vento	float	Wind direction (°)
	Vel_Vento	float	Wind speed (m/s)
Measurement conditions	Data	datetime	Measurement date and hour
	CodNum	ineger	Number of the monitoring station
	Estação	string	Name of the monitoring station
	Lat	float	Latitude position of the station
	Lon	float	Longitude position of the station

Table 1: Measured parameters by the program MonitorAr.

1. Possíveis formas de imputação: estimação polinomial de 2ª ordem. Alguns testes simples pode ser interessante. Para locais onde não há estimação, não faz sentido imputar.

5.2.2 Data transformation

1. Transformação Yeo-Johnson

5.2.3 Feature extraction

From the variable `Data`, we can observe (Figure 1) that 2011 has less observations, because there were only four of the eight stations operating. For that reason, we do not consider the data from this year.

1. Analisar sazonalidade. Adicionar termo seno e cosseno de forma que exista sazonalidade diária, isto é,

$$\text{hour\_sin} = \sin(2\pi \text{ hour}/24)$$

2. Create variable season.

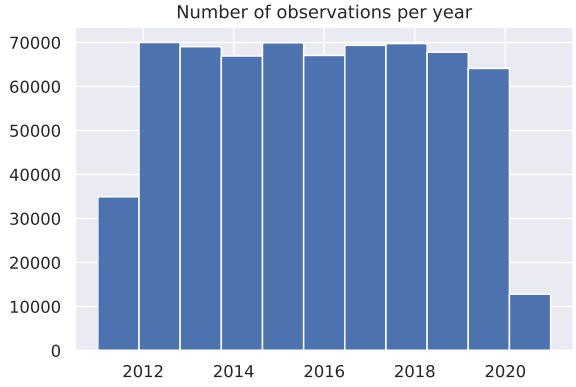

Figure 1: Number of hourly measurements per year. In 2011, only half of the monitoring stations worked.

6 Results

1. Quantitative results of my experiments.

2. Statistical significance.

Monitoring station	**Measured gases/particulates**
Centro (CA)	O_3, CO, PM_{10}
Copacabana (AV)	SO_2, O_3, CO, PM_{10}
São Cristóvão (SC)	SO_2, O_3, CO, PM_{10}
Tijuca (SP)	SO_2, NOx, O_3, CO, PM_{10}
Irajá (IR)	SO_2, NOx, O_3, CO, HC, $PM_{2.5}$, PM_{10}
Bangu (BG)	SO_2, NOx, O_3, CO, HC, PM_{10}
Campo Grande (CG)	SO_2, NOx, O_3, CO, HC, PM_{10}
Pedra de Guaratiba (PG)	O_3, PM_{10}

Table 2: Pollutant data measured by each monitoring station. CO and HC are measured in (ppm), while the others are measured in (µg/m3).

References

[1] da Cidade do Rio de Janeiro, P. (2021). Dados horários do monitoramento da qualidade do ar - monitorar. https://www.data.rio/datasets/PCRJ::dados-hor%C3%A1rios-do-monitoramento-da-qualidade-do-ar-monitorar/about.

[2] Morton, C. (4 de abril 2019). The 50 most beautiful cities in the world. https://www.cntraveler.com/galleries/2016-01-08/the-50-most-beautiful-cities-in-the-world.

[3] Sherrard, M. (2018). Gases that cause air pollution. https://sciencing.com/gases-cause-air-pollution-7445467.html.