Machine Learning

Lucas Machado Moschen

School of Applied Mathematics,
Fundação Getulio Vargas

May 28, 2021

Abstract

Air pollution is one of the contemporary most important topics.

1 Introduction

Air quality is a growing concern and area of research, because most of the cities world-wide have been facing problems with it in the past few decades [3]. The rapid increase of the urban population and the development of the cities is causing environmental pollution, what can give rise to damage on human health.

The emission and transmission of air pollutants, such as, Nitrogen dioxide (NO_2), Carbon monoxide (CO), Ozone (O_3) and Particulate Matter (PM), results in the ambient air pollution, which can be caused by different factors. The World Health Organization (WHO) explained [5] that PM, O_3, and NO_2 have, respectively, the strongest effects on health of air quality.

In Rio de Janeiro city, the city hall recognized the problem and created the Program MonitorAr-Rio in 2008 [2]. The objective was to monitor the air quality in the city and inform the results to the population. Eight fixed stations monitor the main pollutants defined in the legislation, and some meterological conditions.

It is important to have updated knowledge and accurate predictions of the air pollutants, in order to help the formulation of public health and environmental policies. This study proposes models for hourly/daily air quality forecasting for the city of Rio de Janeiro, using the algorithms XXX and XXX.

The text is organized as follows. Section 2 defines the problem clearly and mathematically. Section 3 gives a background on the topic of air quality and air pollution. Section 4 presents the methodology of the work. We present an exploratory data analysis with the data path in Section 5. Section 6 presents the methods used and the experiments related to each one. Sections 7, 8, and 9 presents the results and concludes it.

2 Problem definition

We want to produce time predictions of some pollutants for the Rio de Janeiro city, considering the weather, location, and time variables. We also want to develop a method for estimating the air quality of not monitored regions by the program based on the monitored ones.

Let Y_i be the random variable indicating the quantity of ith pollutant measured in a specific monitoring station, for instance, the quantity of ozone. If $\boldsymbol{X} \in \mathcal{X} \subseteq \mathbb{R}^p$ is the random vector of meterological conditions measured, t is the time of measurement,

and $s_1, ..., s_8$ the positions of the monitoring stations, we have that

$$Y_i = f_i(t, s_k, \boldsymbol{X}) + \epsilon_{t,s_k,i} \quad (1)$$

where $\epsilon_{t,s,i}$ is a random variable with mean zero. After observing, for $t_1 < ... < t_n$,

$$y_i^{t_1,s}, ..., y_i^{t_n,s}$$

and

$$\boldsymbol{x}^{t_1,s}, ..., \boldsymbol{x}^{t_n,s},$$

we want to make predictions $y_i^{\bar{t},s}$ for $\bar{t} > t_{n+1}$. We also want to predict $y_i^{t,\tilde{s}}$ for $\tilde{s}$ different of s_k, for $k = 1, ..., 8$.

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.

Fossil fuels, agriculture, industries, natural disasters are the main causes of air pollution.

3.1 Polluting gases

The atmosphere of the Earth is a dynamic and complex system of natural gases, which are necessary to life [4]. 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

We utilize the Pearson correlation, the mean absolute error (MAE), the root mean squared error (RMSE), and the normalized RMSE (nRMSE) as measures to compare the models. The methods were trained using XXX% of the available data, which correspond to the period XXX. We also forecast of the air quality index (AQI). The software used for performing this experimental phase was developed in Python (version 3.9), mainly using the Pandas and Scikit-learn.

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.

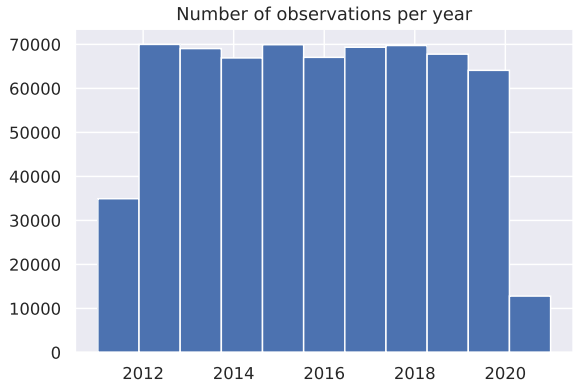

Figure 1: Number of hourly measurements per year. In 2011, only half of the monitoring stations worked.

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.

3. Visualize the series autocorrelation in order to define the number of lag variables per pollutant and particle.

4. Complete the dataset with temporal variables.

	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.

5. The number of features is XXX: lag features, roll mean features, weekend, season, trigonometric

5.2.4 Feature selection

In order to reduce the dimensionality of the feature space, from the XXX variables described, variable selection was performed. Filters are methods which perform feature selection, as well as embedding methods, although they depend on the machine learning algorithm chosen.

1. Pearson correlation-based filter to feature selection. Observe correlation of the lag variables.

The objective of this work is generate different forecasting models for the pollutants A, B, C, and D. The maintained features in the dataset can be seen in Table XXX.

We applied PCA in order to reduce the dimensionality.

6 Experiments with methods

6.1 Method X

The method X has XXX hyperparameters that need to be defined. Time-series split combined with random grid search was used to obtain the optimal numbers. We defined the range to be XXX. The number os iterations chosen to run the random search is XXX. Table XXX shows the results of the random search.

6.2 Method Y

6.3 Method Z

7 Results

1. Quantitative results of my experiments.
2. Statistical significance.

8 Discussion and Future work

9 Conclusion

Métodos estudados no curso

1. Regressão Linear
2. Regressão Logística
3. Análise de discriminante linear
4. KNN
5. Validação cruzada
6. Bootstrap
7. Regularização Ridge e Lasso
8. Árvore de Regressão, Bagging, Random Forest, Boosting
9. Support Vector Machine
10. PCA
11. Métodos de clusterização.
12. Mistura Gaussiana.
13. Redes neurais

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] do Meio-Ambiente, S. (2011-2012). Qualidade do ar na cidade do rio de janeiro: Relatório da rede monitorar-rio. Technical report, Rio de Janeiro.

[3] Mayer, H. (1999). Air pollution in cities. *Atmospheric Environment*, 33(24):4029–4037.

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

[5] WHO (2006). *Air Quality Guidelines: Global Update 2005; Particulate Matter, Ozone, Nitrogen Dioxide and Sulfur Dioxide*. WHO.