Air pollution forecasting in Rio de Janeiro

Final assignment on the subject Machine Learning

Lucas Machado Moschen

School of Applied Mathematics,
Fundação Getulio Vargas

June 7, 2021

Abstract

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

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 [6]. 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 [13] 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 [11]. The objective was to monitor the air quality in the city, in order to verify the degree of exposure to the pollutants of the population, and inform the results to the population. Eight fixed stations monitor the main pollutants defined in the legislation, and some meterological conditions, such as, for example, temperature, relative humidity, solar radiation, and wind.

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 [See Section 3.1 for a detailed description] 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

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

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 predict $y_i^{\bar{t},s}$ for $\bar{t} > t_{n+1}$. We also want to predict $y_i^{t,\tilde{s}}$ for $\tilde{s}$ different from 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. According to Conselho Nacional do Meio Ambiente (CONAMA) [1], atmospheric pollution is

> "... any form of matter or energy with intensity and in quantity, concentration, time or characteristics in disagreement with the established levels, and that make or may make the air: inappropriate, inconvenient, harmful to the environment or harmful to safety."

Different anthropogenic process emit air pollutants, such as, for instance, fossil fuels (motor traffic and domestic), agriculture, and industries. Besides that, natural disaster are an important gas emitter, not controllable, though.

3.1 Polluting gases

The atmosphere of the Earth is a dynamic and complex system of natural gases, which are necessary to life [12]. The planet has a defense mechanism that absorbs part of them. However, high levels of gases concentration can cause several effects in the living beings. Air pollutants are divided according to their origin [13]:

- **Primary:** those emitted into the atmosphere from a source; or
- **Secondary:** those formed within the atmosphere through a chemical reaction.

Other important distinctions are related to their chemical class - organic or inorganic, and related to their physical state - gaseous or particulate. The selected pollutants are:

- **Carbon monoxide (CO):** CO results of incomplete combustion of matter with carbon and it does not present smell or color. Its concentration level is strongly related to car traffic, in addition to agricultural and forest fires. When breathed, it reduces the ability of oxygen transport by the hemoglobins.
- **Ozônio (O_3):** It exists naturally in the atmosphere, where it has the function of absorbing ultraviolet radiation from the sun and of reducing its impact on the planet Earth surface. Nitrogen oxides and organic compounds, with oxygen and high temperatures form the ozone. However, in the troposphere, it is toxic and even explosive. Its effect on health can be drastic.

- **Particulate matter (PM_{10}):** It is composed by particles of solid or liquid matter suspended in the air, with 10 micrometers of less. Combustion is one of the major human sources of particulate matter. Its effects include respiratory tract infections and damages to the environment. When the aerodynamic diameter is between 2.5 and 10 micrometers, we call it PM_10, and it is the most common among the suspended particles measured.

Other measured pollutants that we will not study by computacional and data limitations are Sulfur Dioxide (SO_2), Nitrogen Oxides (NOx), and Hydrocarbons (HC).

3.2 Urban air quality problems

Air pollution episodes, such as, for example, the 1930 Meuse Valley in Belgium [8], the 1952 Great Fog of London [9], and the 2006 Southeast Asian Haze [5] raised questions and concerns about high levels of pollutants in urban ambients, which caused several deaths and hospitalizations. The rapidly expanding populations and cities also increase the exposure to them.

In Brazil, in the last decade, more than 60,000 died on average per year due to air pollution [4]. Household air pollution from solid fuels caused most of the cases, followed by ozone pollution, and particulate matter pollution.

3.3 Air quality index

The air quality index is a synthetic indicator what simplifies the divulgation and the communication among population, private and public sectors, ONGs, etc. Its divulgation is made through diary reports for each monitoring station[1] as shown in Figure 1.

The index considers the pollutants PM_{10}, $PM_{2.5}$, O_3, CO, NO_2, and SO_2. For each one, The AQIr is calculated as follows,

$$AQIr = I_{ini} + \frac{I_{fin} - I_{ini}}{C_{fin} - C_{ini}} \times (C - C_{ini}), \qquad (2)$$

such that I_{ini} (I_{fin}) is the value of the index that corresponds to the initial (final) concentration of the range; C_{ini} (C_{fin}) is the initial (final) concentration of the range in which the measured concentration is located, and C is the concentration measured. The value informed is the maximum $AQIr$ of the pollutants.

After calculating the AIQ, we classify air quality at five levels: N1 (0-40), N2 (41-80), N3 (81-120), N4 (121-200), and N5 ($>$ 200). For more details, consult the technical guide from the Ministry of the Environment [7]. In this work, we only consider the pollutants described on Section 3.1 for the forecasting of the air quality index.

4 Methodology

Exploratory data analysis is the first step into the project, after identifying the problem. We use visualization and descriptive statistics to summarize the most important information to have insights. After this, we make data preprocessing, what includes missing data imputation, outlies, and feature engineering.

We utilize 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 70% of the available data, considering the first years. 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-Rio [10]. 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 per monitoring station. The events were collected between

[1] http://jeap.rio.rj.gov.br/je-metinfosmac/boletim

MonitorAR Rio
Programa de Monitoramento da Qualidade do Ar

BOLETIM DE QUALIDADE DO AR

01/03/2018 15:00H
Quinta-feira

Exibir Boletim Anterior:
Data: Exibir

Rio PREFEITURA
Meio Ambiente da Cidade

Estação	Concentração Máxima Poluentes Monitorados					Índice de Qualidade do Ar (IQA)	Classificação
	Material Particulado (MP_{10}) [$\mu g/m^3$]	Ozônio (O_3) [$\mu g/m^3$]	Monóxido de Carbono (CO) [ppm]	Dióxido de Nitrogênio (NO_2) [$\mu g/m^3$]	Dióxido de Enxofre (SO_2) [$\mu g/m^3$]		
Centro	25,8	72,9	ND	NM	NM	46	Boa
Copacabana	46,6	ND	0,1	NM	1,6	47	Boa
São Cristóvão	19,7	84,5	0,1	NM	8,1	53	Regular
Tijuca	26,3	71,9	0,6	65,9	2,4	45	Boa
Irajá	29,0	66,3	0,4	ND	ND	41	Boa
Bangu	37,6	126,3	0,7	44,8	3,9	79	Regular
Campo Grande	30,2	115,9	0,6	52,2	ND	73	Regular
Pedra de Guaratiba	34,6	75,2	NM	NM	NM	47	Boa
Unidade Móvel Recreio	47,1	82,6	0,2	NM	ND	52	Regular

Condições Meteorológicas observadas no período:

A atuação de áreas de instabilidade ocasionou predomínio de céu parcialmente nublado, porém sem registros de chuva sobre a Cidade. Desta forma, observou-se a manutenção das concentrações dos poluentes em relação ao dia anterior, em que a qualidade do ar ficou classificada como BOA e REGULAR nos locais monitorados.

Tendência da Qualidade do Ar para as Próximas 24h:

O novo posicionamento do sistema de alta pressão, associado a áreas de instabilidade favorecerão nebulosidade variada com possibilidade de chuva fraca em áreas isoladas. Assim, espera-se a manutenção das concentrações dos poluentes, o que deverá deixar a qualidade do ar classificada como BOA e REGULAR nas localidades monitoradas.

Figure 1: Report from March 1^{st}, 2018.

January 1, 2011, and March 31, 2021. A total of 661,662 records were used.

We count the missing values for each of the main variables used in this work and present them in Table 3. We observe that some meterological variables have more than 10% of missing values, what is a lot. The CO gas has more missing values than the other gases, because Pedra de Guaratiba does not measure it. If we disconsider this kind of absent values, CO has around 6%. Imputation methods to handle this problem in Section 5.2.

It is also noticed that almost 91% of the values in `Chuva` column and 26% of `RS` column are zero. If we consider the accumulated monthly amount of rain, the value seems to make sense and it is comparable to other sources, therefore we kept it.

We did not observe any pattern of missing values per year or per monitoring station. In 2012, over half of the wind information is missing, however other features are stable. However, after 2016, UR dominates the number of absent data. If we aggregate by time (hour, day, month, and year), and sum up the values of the features of each station, we observe little missing data. This implies we can use information of other stations to impute data whenever necessary.

We see the summary statistics of the variables in Table 4. The high skewness and kurtosis from `Chuva` column reaffirm we have heavy tails, or outliers. We can say the same for `Press`. This is interesting because, it is common to see days with extreme values in meterological variables. In particular, the only hours with precipitation greater than 100 mm were in May 2020 in Tijuca, what is confirmed by weather references [2]. CO and PM_{10} have high kurtosis also, so we shall give an attention for outliers too.

The time series of the three gases are presented in Figure 2. We observe an increase in ozone levels during the year and a season effect, what corroborates on the way ozone is formed. It also seems there is a reduction in variability in CO and PM_{10} levels. The year of 2020 show a reduction in CO apparently, what is explained by the Coronavirus pandemic.

We confirm the tendencies of PM_{10} and O_3 in Figure 3. It is important to note that 2021 is not finished, so season effects are not complete yet.

By Figure 4, years 2011 and 2021 have less observation than the others. The year 2021 did not end as previously mentioned and the year 2011 had less

	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	integer	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.

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).

Variable	Missing values
Chuva	15812 (2.38 %)
Pres	15294 (2.31 %)
RS	48260 (7.29 %)
Temp	70617 (10.6 %)
UR	110619 (16.7 %)
Dir_Vento	90498 (13.6 %)
Vel_Vento	90743 (13.7 %)
CO	114179 (17.2 %)
O3	37133 (5.61 %)
PM10	36142 (5.46 %)

Table 3: Missing data in absolute and proportional values of the main variables. Data, CodNum, Lat, and Lon do not have nan values.

monitoring stations.

Figure 5 shows an overview of the correlations between different features of the data to identify possible linear relations. The scatter plots of two to two features represents it with more details, but there much data, and the image has a large size. It does not present anything much different, though. The variables `Temp`, `UR`, and `RS` are strongly linearly related with absolute correlation greater than 0.6.

We also observe an interesting behavior of the ozone during the day, as showed in Figure 6. For each hour of the day, we drew a boxplot to represent the distribution of ozone in the respective hour. We observe that (1) the pollutants have heavy tail, or this data has a lot of outliers; and (2) the medians go along with the movement of the sun.

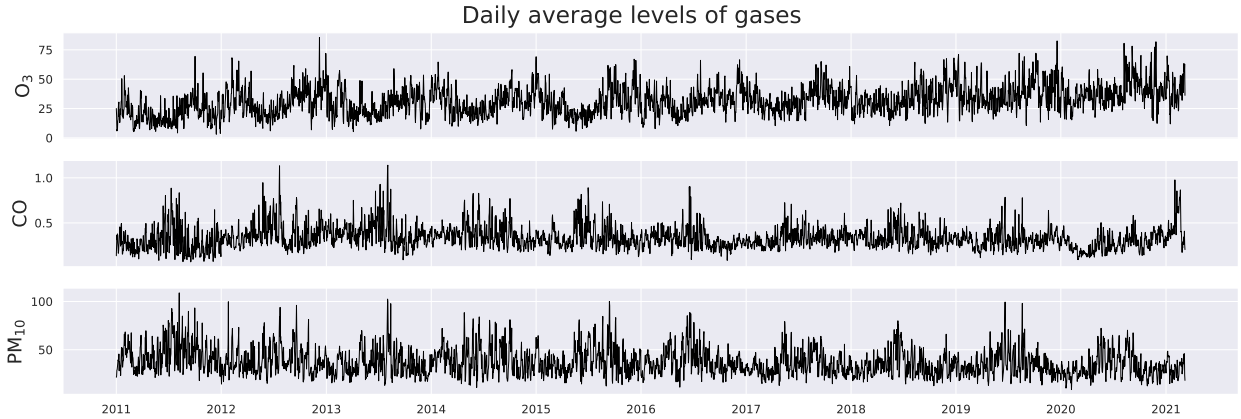

Figure 2: Time series with diary average levels.

	Chuva	Pres	RS	Temp	UR	Dir	Vel	CO	O3	PM10
Mean	0.13	1014.65	152.82	26.12	70.90	163.73	1.21	0.34	31.98	36.91
Std	1.64	5.68	244.37	4.90	18.35	73.45	1.00	0.28	29.81	23.52
Min	0.00	800.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	0.00	1011.12	0.00	22.67	58.39	100.00	0.55	0.14	8.68	21.00
50%	0.00	1014.30	6.17	25.54	72.75	166.17	0.92	0.29	24.52	32.00
75%	0.00	1018.02	224.00	28.99	85.08	222.50	1.55	0.46	46.89	47.00
Max	426.60	1036.48	1864.67	49.08	100.00	358.83	25.50	12.08	355.45	994.00
Skew	114.55	-7.32	1.61	0.55	-0.44	0.04	3.74	2.75	1.56	2.72
Kurt	23177.40	282.90	1.48	0.33	-0.40	-0.97	47.30	24.85	3.71	38.67

Table 4: Statistics of the meterological variables and gases.

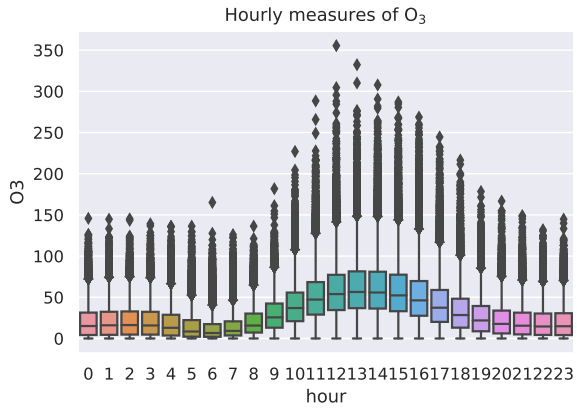

Figure 6: Boxplot of hourly ozone measurements for each hour. It seems to accompany the sun during the day.

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.

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.

1. Possíveis formas de imputação: estimação polinomial de 2ª ordem. Alguns testes simples pode

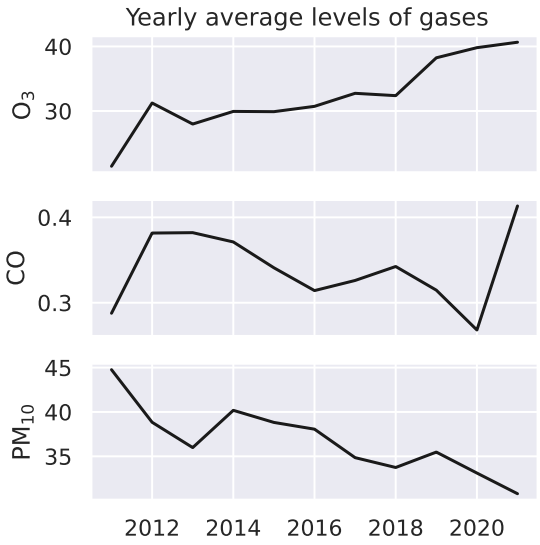

Figure 3: Time series with yearly average levels.

ser interessante. Para locais onde não há estimação, não faz sentido imputar.

5.2.2 Data transformation

We applied the Yeo-Johnson power transform [14] on continuous variables in order to approximate the data distribution to a Gaussian distribution and to decrease the heteroscedasticity, as suggested by [3]. In figure 7, the gases distribution (disconsidered missing data imputation) are shown before and after the transformation. The selected λ for each feature was estimated through maximum likelihood. Following the order from table 4, the values were, approximately, -21.70, 11.99, -0.11, 0.28, 1.53, 0.86, -0.81, -1.58, 0.25, and 0.27.

5.2.3 Feature extraction

From the variable `Data`, we could extract the variables `year`, `month`, `day`, and `hour`. We also create a boolean variable for the weekend, and a `season` variable with values 0, 1, 2, or 4. In order to consider

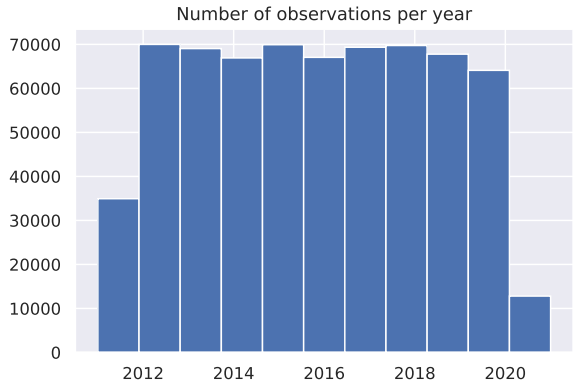

Figure 4: Number of hourly measurements per year. In 2011, only half of the monitoring stations worked.

the hourly seasonality, we also add the variables

$$\text{hour\_sin} = \sin(2\pi \text{ hour}/24),$$
$$\text{hour\_cos} = \cos(2\pi \text{ hour}/24).$$

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

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

5.2.4 Feature selection

Variables to remove: `hour` since it is highly correlated to `hour_sin`, and `Data`, because it has no service anymore. We will have only numeric variables from now on. We need to observe correlation between lag variables from the pollutants

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 num-

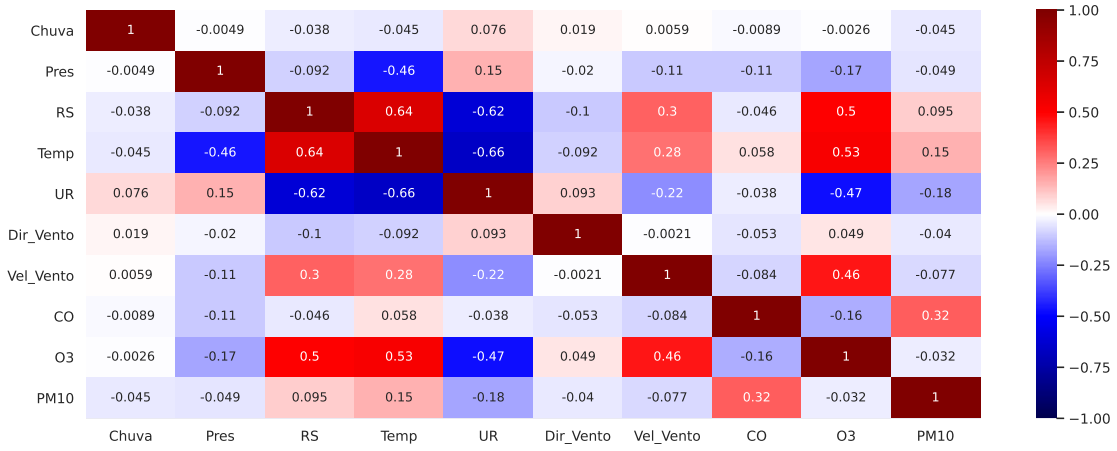

Figure 5: Correlation heat map comparing the data features.

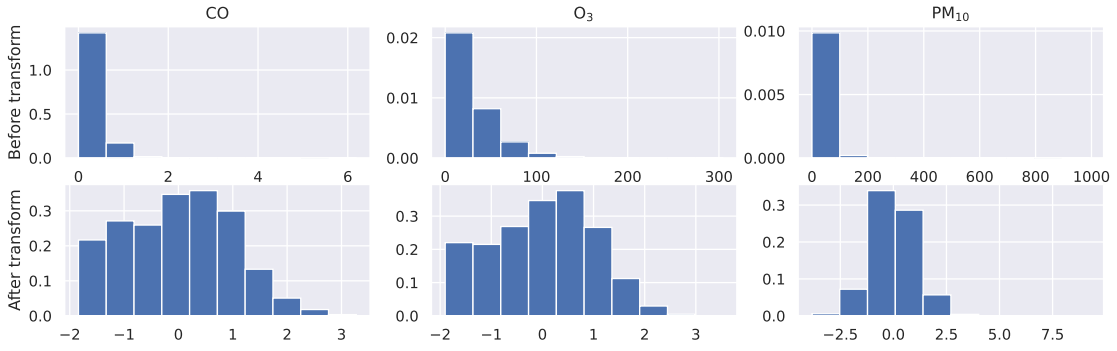

Figure 7: Gases distribution before and after the power transform.

bers. 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

References

[1] BRASIL. Conselho Nacional do Meio Ambiente (1990). Resolução conama n° 003, de 22 de agosto de 1990. Available at https://www.ufjf.br/baccan/files/2012/11/Resolucao_003_CONAMA_de_1990-Padroes-de-qualidade-do-ar.pdf.

[2] Climatempo (2020). Média de chuva de maio no rio de janeiro fica acima do normal. Available at https://www.climatempo.com.br/noticia/2020/06/01/media-de-chuva-de-maio-no-rio-de-janeiro-fica-cima-do-normal-3857.

[3] Gocheva-Ilieva, S. G., Ivanov, A. V., Voynikova, D. S., and Boyadzhiev, D. T. (2014). Time series analysis and forecasting for air pollution in small urban area: an sarima and factor analysis approach. *Stochastic environmental research and risk assessment*, 28(4):1045–1060.

[4] Health Effects Institute (IHME, 2020). State of Global Air 2020. Data source: Global Burden of Disease Study 2019. Available at https://www.stateofglobalair.org/data/#/health/plot.

[5] Jones, D. S. (2006). Asean and transboundary haze pollution in southeast asia. *Asia Europe Journal*, 4(3):431–446.

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

[7] Ministério do Meio Ambiente (2019). Guia técnico para o monitoramento e avaliação da qualidade do ar. Available at https://www.gov.br/mma/pt-br/centrais-de-conteudo/mma-guia-tecnico-qualidade-do-ar-pdf.

[8] Nemery, B., Hoet, P. H., and Nemmar, A. (2001). The meuse valley fog of 1930: an air pollution disaster. *The lancet*, 357(9257):704–708.

[9] Polivka, B. J. (2018). The great london smog of 1952. *AJN The American Journal of Nursing*, 118(4).

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

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

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

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

[14] Yeo, I.-K. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959.