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 15, 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), result 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 of the population to the pollutants, 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 air quality forecasting for the city of Rio de Janeiro.

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. Section 5 contains a description of the data

used in this work and an exploratory data analysis with the data path. Section 6 presents the methods used and the experiments related to each one. Sections 7, 8, and 9 end the text with main results and conclusions.

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, then it is measured

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

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 $\bar{t} > t_{n+1}$ and $y_i^{t,\tilde{s}}$ for $\tilde{s} \notin \{s_k : k = 1, ..., 8\}$.

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, own translation], 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 processes emit air pollutants, such as, for instance, fossil fuels (motor traffic and domestic), agriculture, and industries. Besides that, natural disasters 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 for this work are:

- **Carbon monoxide (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, the name is 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 to the population, private and public sectors, ONGs, among others. Its divulgation is made through diary reports for each monitoring station[1] as shown in Figure 1.

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

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

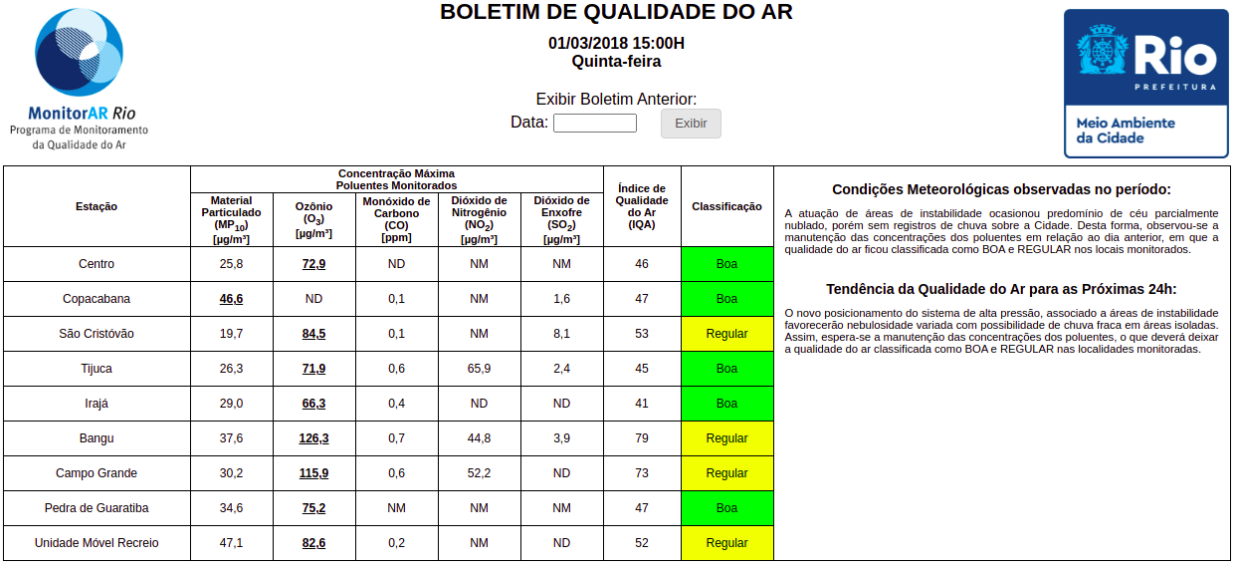
Exploratory data analysis is the first step into the project, after identifying the problem. The proposed EDA includes visualization and descriptive statistics to summarize the most relevant information to have insights. After this, we make data preprocessing, which includes missing data imputation, outlies, and feature engineering.

The evaluation methods for the algorithms are the mean absolute error (MAE), the root mean squared error (RMSE), and the normalized RMSE (nRMSE). 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 con-



BOLETIM DE QUALIDADE DO AR

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

Exibir Boletim Anterior:
Data: Exibir

MonitorAR Rio
Programa de Monitoramento da Qualidade do Ar

Rio PREFEITURA
Meio Ambiente da Cidade

Estação	Concentração Máxima Poluentes Monitorados					Índice de Qualidade do Ar (IQA)	Classificação
	Material Particulado (MP_{10}) [µg/m³]	Ozônio (O_3) [µg/m³]	Monóxido de Carbono (CO) [ppm]	Dióxido de Nitrogênio (NO_2) [µg/m³]	Dióxido de Enxofre (SO_2) [µg/m³]		
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.

tains 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 start on January 1, 2011, and end on March 31, 2021, totalizing 661,662 records.

The missing values (Table 3) for each of the main variables used in this work are around 10%. Some meteorological variables have more than 10% of missing values, which is a lot. The CO gas has more missing values than the other gases because Pedra de Guaratiba does not measure it. When this kind of absent value is disregarded, CO has around 6%. Imputation methods in Section 5.2.

Almost 91% of the values in the `Chuva` column and 26% of the `RS` column are zero. If we consider the accumulated monthly amount of rain, it seems to make sense, and it is comparable to other sources.

We did not observe any pattern of missing values per year or 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, there are less missing data, proportionally, what. This implies other stations can provide useful information.

Table 4 contains the summary statistics of the variables. The high skewness and kurtosis from `Chuva` column reaffirm the heavy tails of its distribution, or the presence of outliers. The same occurs with `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 and, therefore, can have outliers.

The time series of the three gases are presented in Figure 2. It indicates an increase in ozone levels during the years 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 fin-

	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.

ished, 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 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.

Figure 6 shows an interesting behavior of the ozone during the day. For each hour of the day, we represent the distribution of ozone in the respective hour. We

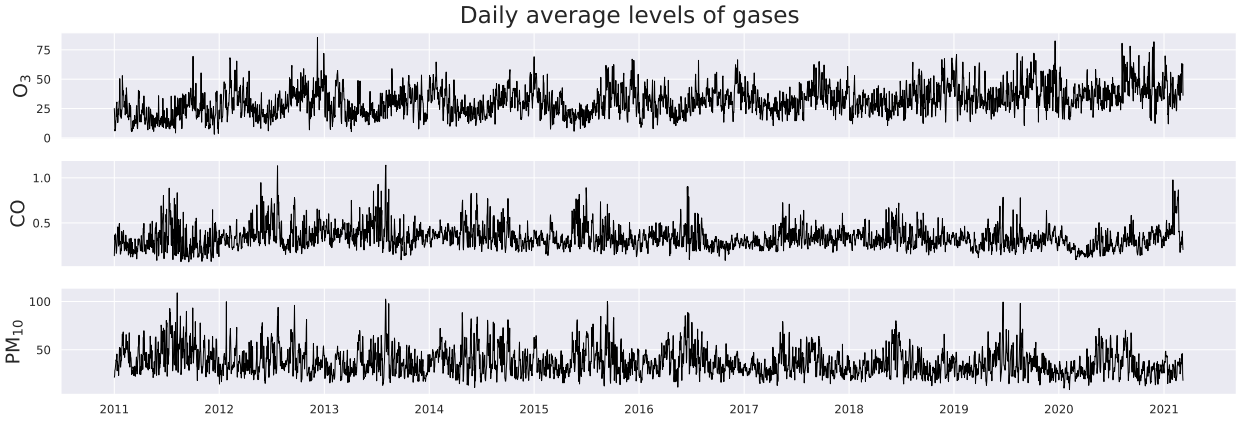


Figure 2: Time series with diary average levels.

	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.

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.

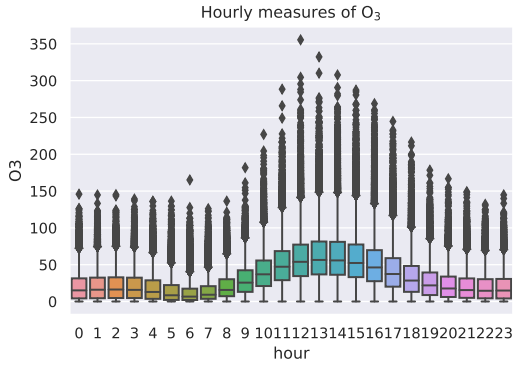


Figure 6: Boxplot of hourly ozone measurements.

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. Data from year 2020 and 2021 will be removed given the pandemic in the world.

5.2.1 Seasonal and time features

From the variable `Data`, it is extracted the variables `year`, `month`, `day`, `hour`, a boolean variable indicating the weekend, and a `season` variable. In order to consider the hourly seasonality, it is created the variables hour_sin $= \sin(2\pi\ \text{hour}/24)$ and hour_cos $= \cos(2\pi\ \text{hour}/24)$.

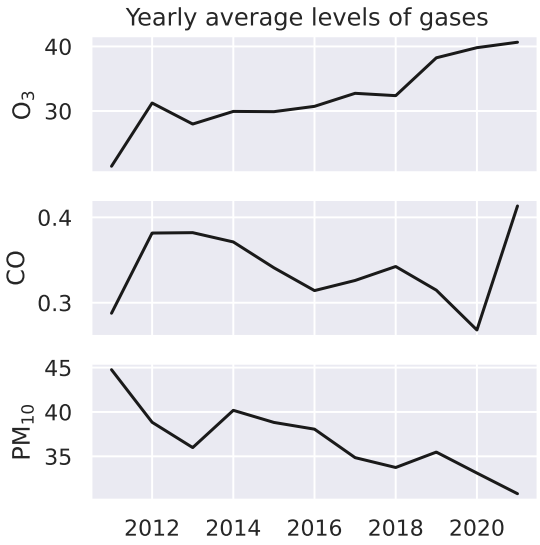


Figure 3: Time series with yearly average levels.

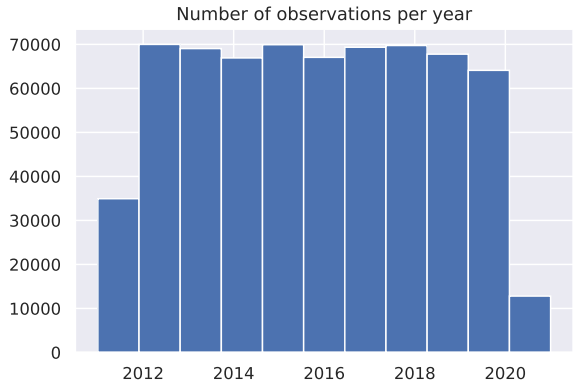


Figure 4: Number of hourly measurements per year. In 2011, only half of the monitoring stations worked.

5.2.2 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 CO in Pedra de Guaratiba; and (2) monitoring stations did not measure in a period for some reason. For the first case, missing values remain in the dataset and the prediction is not performed. For the second case, three methods were compared:

1. **Simple mean imputation:** For each year and each column, the algorithm imputes the mean in the NaN values.

2. **Location:** For each time period, the missing values are replaced by the average among the others measuring stations. For example, if CO is missing at 6h on 01/01/2011 at Centro station, it is replaced by the mean among the other stations in the same time.

3. **k-NN:** For each year, if a feature is missing in row i, the algorithm selects the k closer points according to the nan euclidean measure. This measure calculates the euclidean distance among non NaN entries and rescale depending on the number of them. The years are separated to reduce the number of rows.

The data values are normalized to the range of $[0, 1]$, because k-NN is based on a distance measure. To evaluate these methods, we develop a sample strategy similar to the Bootstrap method. In each simulation, we sample 20% of rows with non NaN values and randomly choose 7% of the cells to be removed (this value was chosen because this was observed across the entire dataset). From the simulated dataset with missing values, the methods impute as explained before. We compare the imputed sample after imputation with the sample before removal and calculate the mean squared error. The results for the year 2016 are above 5.

Método	**MSE**
5-NN	6.504e-04
10-NN	6.504e-04
30-NN	6.504e-04
50-NN	6.504e-04
100-NN	6.504e-04
Location	1.077e-03
Simple Imputation	1.636e-03

Table 5: Results from the imputation

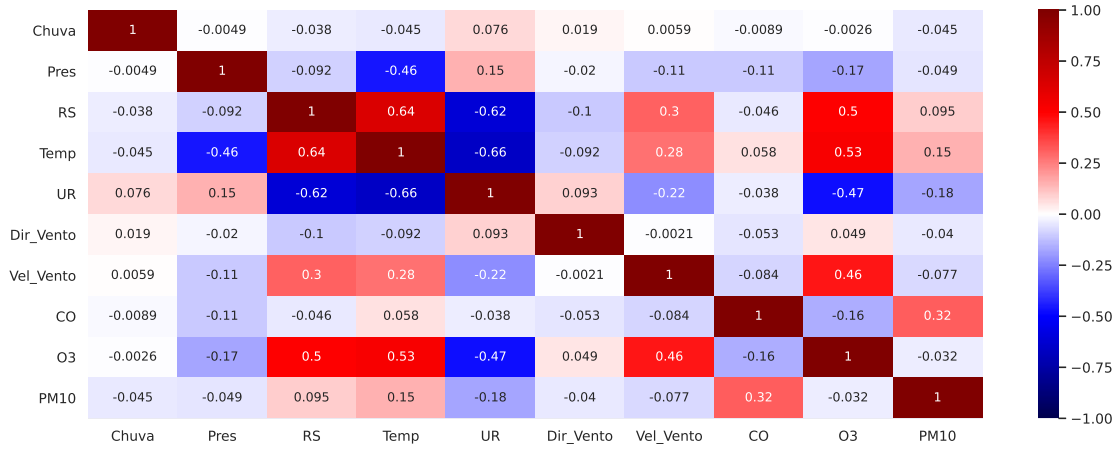


Figure 5: Correlation heat map comparing the data features.

For that reason, we apply the 5-NN in the dataset per year to impute data.

5.2.3 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 (disregarded 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, -20.16, 12.58, -0.1, 0.29, 1.59, 0.81, -0.79, -1.69, 0.28, and 0.27.

5.2.4 Feature extraction

The lag is a time gap in the series and are useful to analyse seasonality. In general, besides the influence of another variables, the past values can be helpful to make good predictions. An exploratory study with autocorrelation (ACF) and partial autocorrelation (PACF) to define the number of lag variables. Figure 8 shows interesting patterns: CO has spikes at 12 multiples. That indicates a high correlation with twelve hours difference. The ozone does not have this quality, but presents the diary spikes. It is interesting that at Bangu station, this is negatively related. Since the PACF has only two spikes in all graphs, there is an autoregressive term in the pollutants series of order two. We add two lag variables for each pollutant and each monitoring station, then. We also add a 24 lag time.

Variables to remove: `hour` since it is highly correlated with `hour_sin`, and `Data`, because it has no service anymore. We will have only numeric variables from now on. The total number of features after all the above processes is 29 including meteorological conditions, time-related, lags, and pollutants measurements.

Given the high dimensionality (26 independent variables), a principal component analysis (PCA) was applied to compare the results when the whole dataset is used.

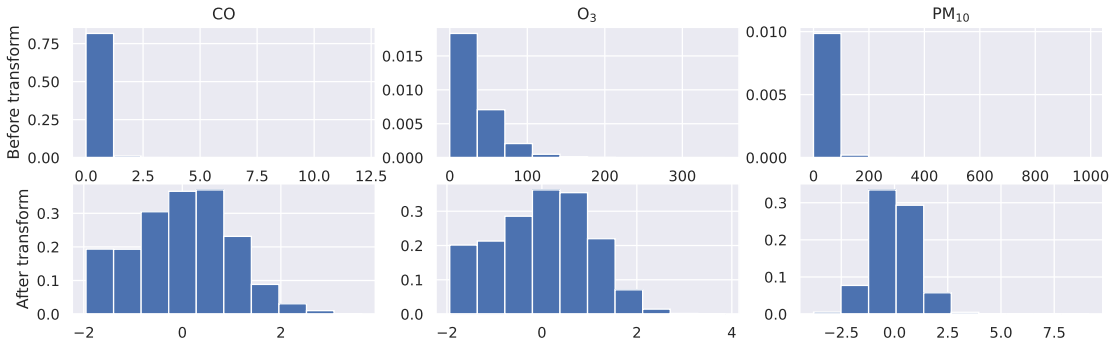


Figure 7: Gases distribution before and after the power transform.

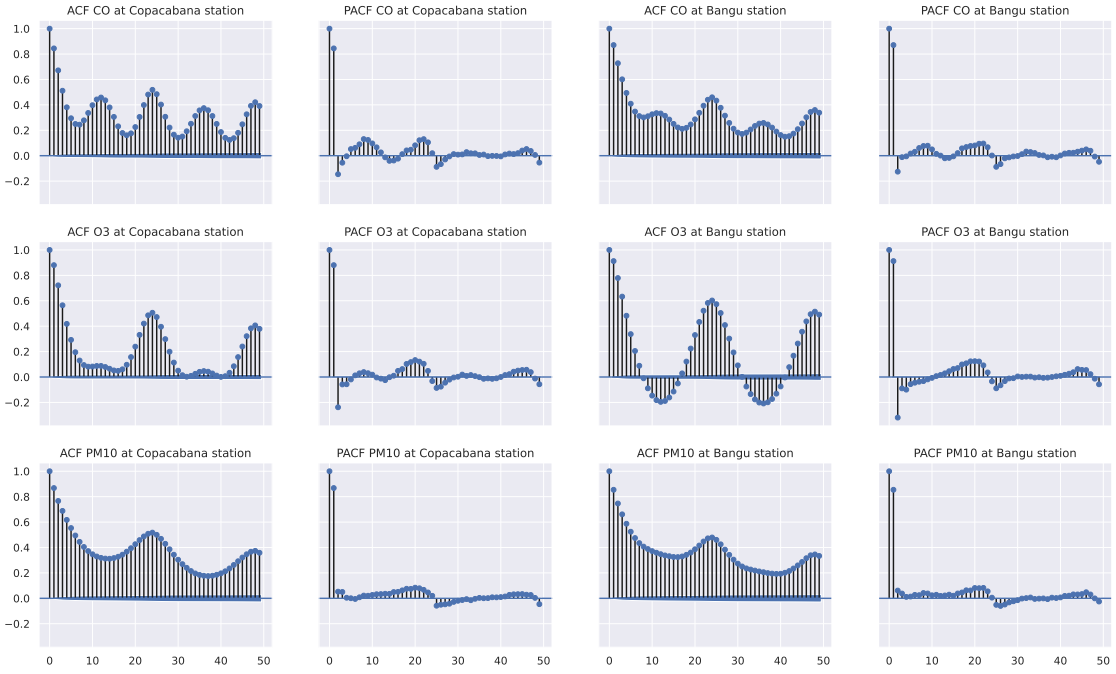


Figure 8: ACF and PACF plots for CO, O_3, and PM_{10} at Copacabana and Bangu stations.

6 Methods and experiment settings

6.1 Linear Regression

6.2 Support Vector Machine

6.3 Random Forest

6.4 Boosting

6.5 Linear Regression + Expectation Maximization

7 Results

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

Lembrar de

1. Testar estacionaridade de cada série;
2. Lembrar de inverter os dados pelo power transformation: ler p2, fazer power transform.

8 Discussion and Future work

9 Conclusion

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.