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 21, 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 [10]. 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 [22] 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 [19]. 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 i^{th} 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 [20]. 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 [22]:

- **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 [12], the 1952 Great Fog of London [16], and the 2006 Southeast Asian Haze [7] 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 [6]. 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 [11]. In this work, we only consider the pollutants described on Section 3.1 for the forecasting of the air quality index.

4 Methodology

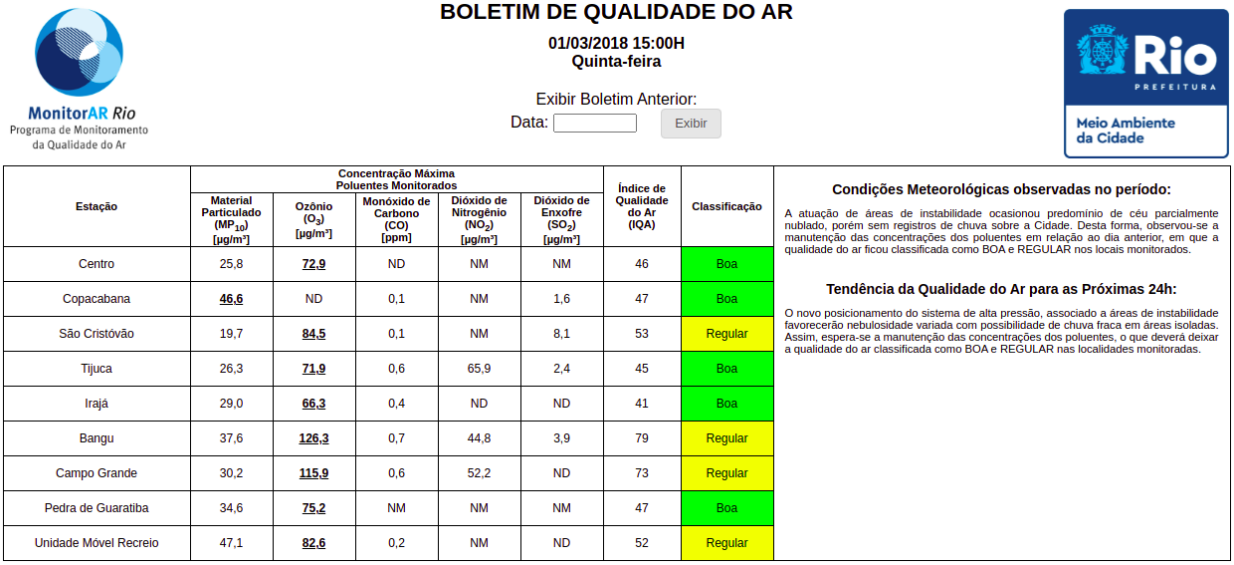
Exploratory data analysis is the first step into the project, after identifying the problem. The proposed EDA includes visualization and descriptive statistics to summarize the most relevant information to have insights. After this, we make data preprocessing, which includes missing data imputation,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

The evaluation methods for the algorithms are the mean absolute error (MAE), the root mean squared error (RMSE), and the rsquared (R^2). 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 (v. 3.9), mainly using the Pandas [13, 21], Numpy [5], Scikit-learn [15], and Statsmodels [18].

5 Exploratory data analysis

5.1 Data description

The dataset used in this study was extracted from the project MonitorAr-Rio [17]. 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

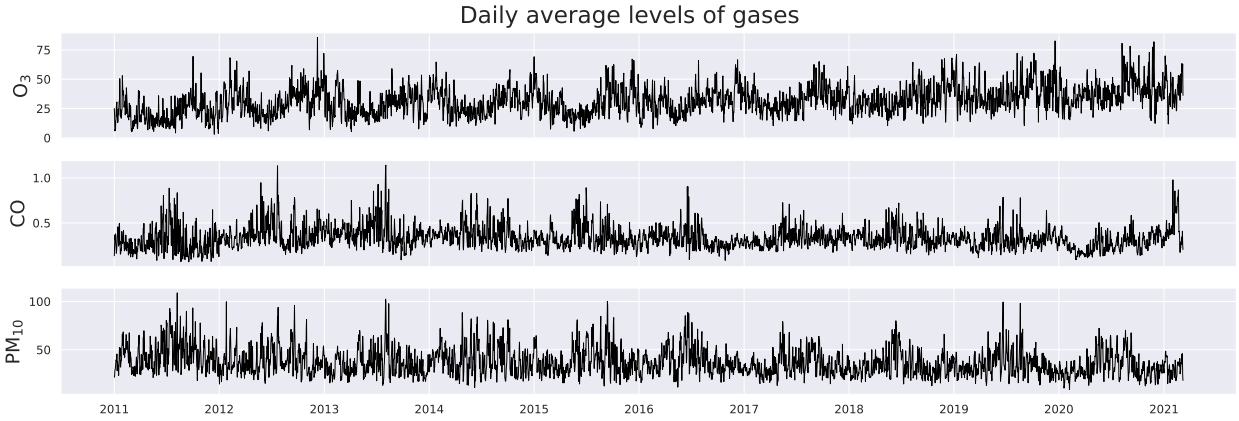


Figure 2: Time series with diary average levels.

	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.

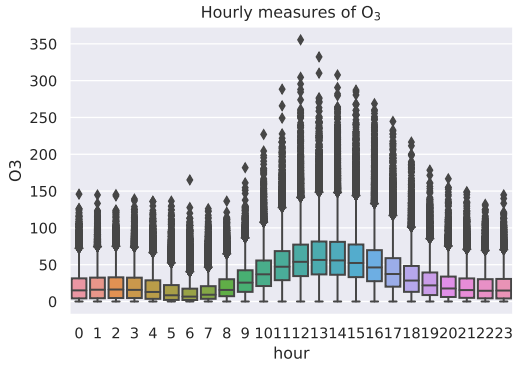


Figure 6: Boxplot of hourly ozone measurements.

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 $= \cos(2\pi\ \text{hour}/24)$.

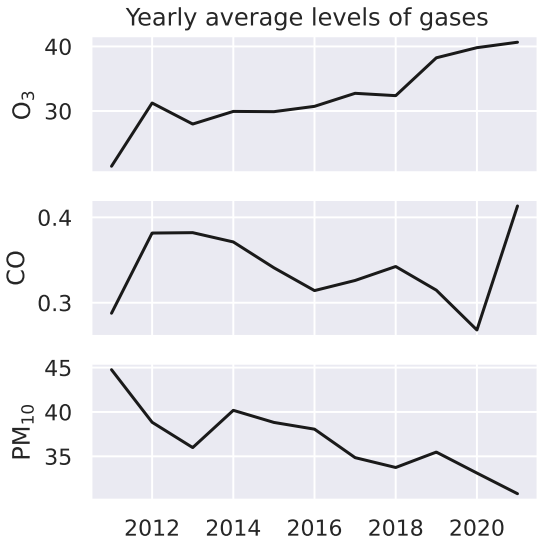


Figure 3: Time series with yearly average levels.

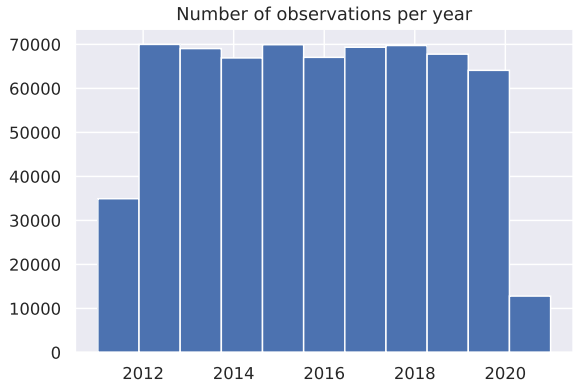


Figure 4: Number of hourly measurements per year. In 2011, only half of the monitoring stations worked.

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

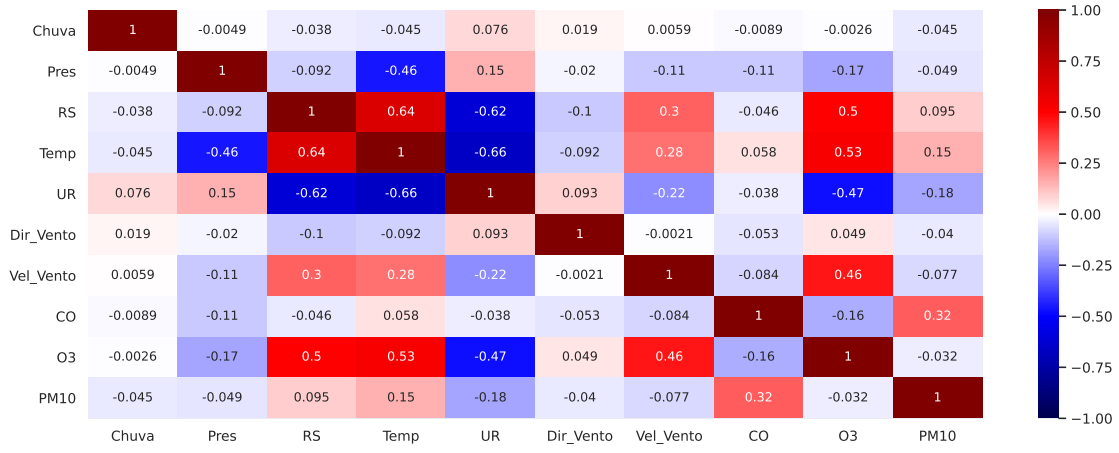


Figure 5: Correlation heat map comparing the data features.

For that reason, we apply the 5-NN in the dataset per year to impute data. It is important to note that the computation was a real barrier for a more deep analysis.

5.2.3 Data transformation

We applied the Yeo-Johnson power transform [23] on continuous variables in order to approximate the data distribution to a Gaussian distribution and to decrease the heteroscedasticity, as suggested by [4]. 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 to catch the diary influence. Besides that, a rolling mean variable with a 24 lag is added to simulate a moving average (MA) term. It is not possible to use the MA model in this framework since they are not observed [9].

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 32 including meteorological conditions, time-related, lags, moving averages, and pollutants measurements.

Given the high dimensionality (27 independent variables), a principal component analysis (PCA) was applied to compare the results when the whole dataset is used. This is done only to verify if all the variables really help the forecasting.

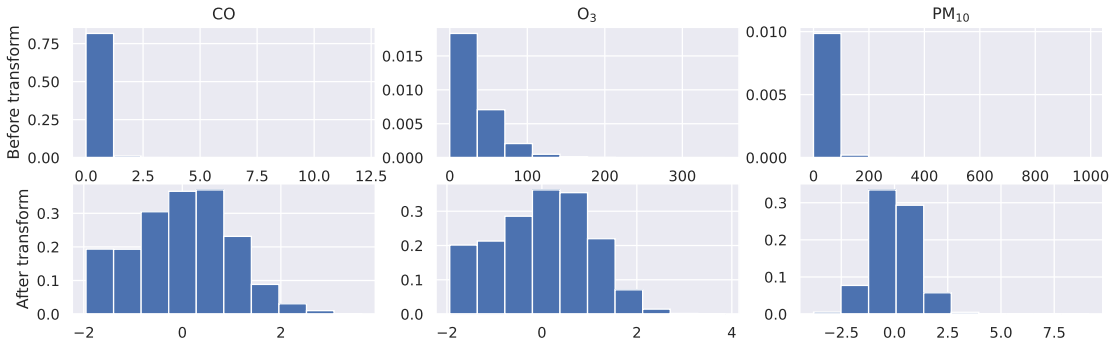


Figure 7: Gases distribution before and after the power transform.

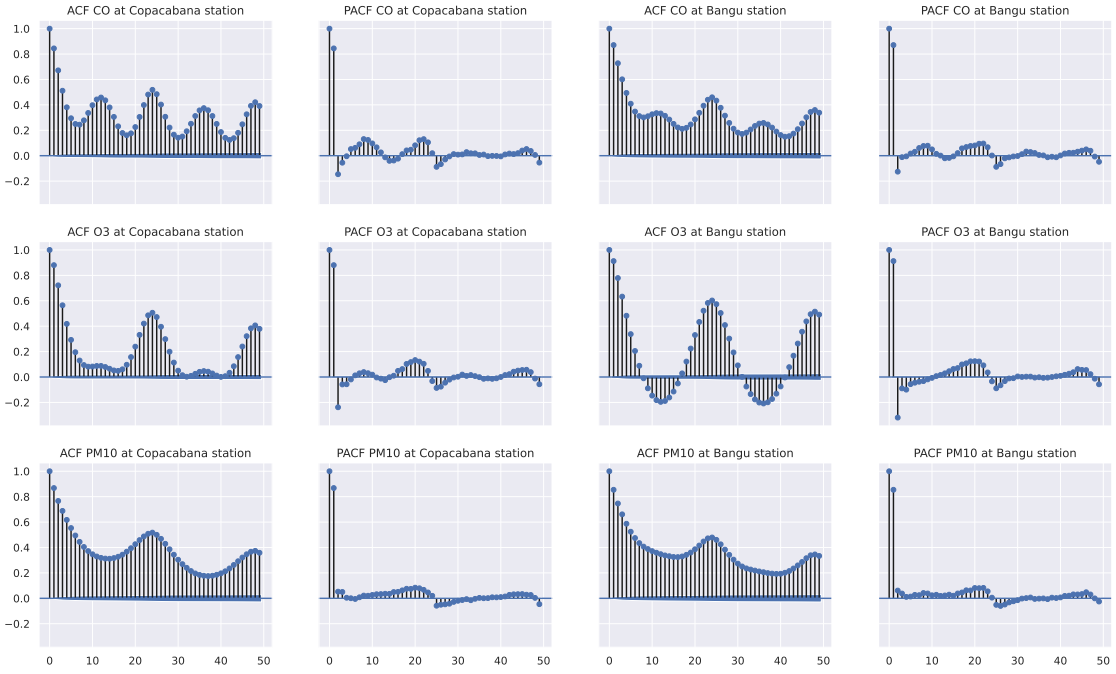


Figure 8: ACF and PACF plots for CO, O_3, and PM_{10} at Copacabana and Bangu stations.

6 Experiment settings

We have already separated the dataset into training and testing. In this section we describe the methods used for the estimation: *Linear Regression*, *Support Vector Machine*, *Random Forest*, *Boosting*, and a model that handle missing data simultaneously with the fitting, using the *Expectation Maximization*. Each gas and each localization has its own model.

6.1 Linear Regression

The first attempt was to consider linear regression. In this case, the model supposes that the expected value of the gases quantity (conditioned on the data) is a linear combination of the independent variables.

First we apply the simple OLS on the whole dataset. In general this is not a great model, because it adds variance on the estimation and, for that reason, we consider a regularization term (elastic net) with two parameters: α to control the penalty, and w_{l1} to control the weight given to $\mathcal{L}_1$ penalty. The parameters are chosen with cross validation (5-Fold). This approach allows adding polynomial and interaction terms, since the $\mathcal{L}_1$ penalty sets several parameters to zero. The 5-Fold is realized with $w_{l1} \in \{0.1, 0.2, ..., 0.9, 1.0\}$ and $\alpha \in \{1, 2, 3, ..., 20\}$.

Other interesting approach to do feature selection, rather than $\mathcal{L}_1$ penalty is Forward Feature Selector. It is slower than Elastic Net, but it selects a best subset of features given that it adds features in a greedy fashion. The estimator chooses the best feature based on cross-validation score (with r2 score) and it stops if the improving in the score is lesser than a threshold.

6.2 Support Vector Regression

It is an extension of Support Vector Machine (SVM) algorithms for regression. Given the more than quadratic complexity of the algorithm, this does not scale for datasets with more than 10,000 samples, as suggested by Scikit-learn User Guide [8, 15]. For that reason, we suppose a linear kernel. The loss function considered is $\mathcal{L}_1$ loss, with parameter ϵ. A regularization parameter C is also added to the model, such that, C is inversely proportional to the strength of the regularization. All columns are scaled to have mean 0 and variance 1. The parameters are calibrated with cross validation (5-Fold), with $C \in \{10^{-3}, 10^{-2}, ..., 10^{2}\}$ and $\epsilon = \{0.001, 0.01, 0.1, 0.2, 0.3\}$. The problem with that approach is that there will be 150 fittings + 150 predictions. If we set the maximum number of iterations to be 5000, the program does not converge, but each of the fitting takes 1min, what is impractical. For that reason, this model is very problematic.

6.3 Random Forest

The random forest regressor is an extension of decision tree with B bootstrap samples such that each split considers m predictors, that is the root of the number of predictors. The parameter c measures the complexity parameter (minimal cost-complexity pruning). The criterion to measure the quality of a split is MSE, in order to reduce variance. The minimum number of samples required to split is the parameter s.

6.4 Linear Regression + Expectation Maximization

In this scenario, we follow the approach developed by [3] and demonstrated by [14]. This method supposes the data comes from a normal distribution with mean $\mu_{y,X}$ and covariance matrix $\Sigma_{y,X}$, including the dependent and independent variables. It uses the Expectation Maximization (EM) algorithm to estimate these parameters, despite the missing data. With the normal parameters estimated, the following formula allows the specification of the regressor parameters:

$$\beta = (\mu_y - \Sigma_{y,X}\Sigma_X^{-1}\mu_X, \Sigma_{y,X}\Sigma_X^{-1})^T.$$

A forward variable selection in the same terms as before is applied. The data transformations 5.2.3 are done without the missing data imputation.

6.5 Summary

Therefore the considered models and hyperparameters are the following:

1. Simple linear regression: all predictors, no hyperparameter.

2. Elastic-net regression: all predictors, α measuring the penalty strength, and w_{l1} the weight for $\mathcal{L}_1$ penalty.

3. Forward Feature Selection + Linear regression: The number of features to select as function of a threshold.

4. Support Vector Regression: all predictors, ϵ measures the loss, and C the regularization term. The variables are transformed between 0 and 1.

5. Random Forest: all predictors, B bootstrap samples, c is the complexity parameter, and s is the minimum number of samples to split.

6. Linear regression + missing data imputation: no additional parameters.

7 Results

The results are separated by gas, but not by monitoring station. We present more detailed results for only one (Tijuca station was chosen randomly), and after aggregated results considering all the stations. This is done because we have more than a hundred models to analyse. To deal with this diversity, the steps (hyperparameters choice and evaluation) are automated. We are making predictions one hour ahead and it is possible to compare with one day ahead models. After this, we will have a best model for each gas and each station (23 on total).

7.1 Tijuca monitoring station

The results follow the order specified in the above summary for each pollutant.

7.1.1 O_3

Simple linear regression

Applying the simple linear regression, the R^2 in the testing set was 0.837, what appears to be a good start fitting. The variable with greater t-statistic was

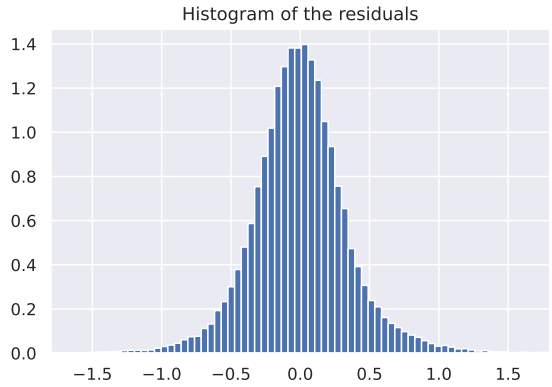


Figure 9: Histogram of the residuals of simple linear regression model.

O_3 shifted by one hour. Other features with great t-statistic (more than 20) was, in order, wind speedy, ozone lag 2, hour sin, UR, ozone lag 24, hour cos, and RS. This is interesting because we have already observed the hourly seasonality and by the formation of ozone, RS is expected to influence (not necessarily linearly). The shifts were expected by the autocorrelation graph. Only some few features (3) had p-value greater that 0.05. The F-statistic considering all variables was practically zero. One important problem with this approach was the very big condition number (3.4e+06) given the observed multicollinearity. Figure 9 shows the histogram of the residuals very similar to a normal distribution (as assumed by the model).The kurtosis was nearly 3, while the skewness around 0.22. Jarque-Bera rejects the null hypotheses that skew and kurt are the same as normal distribution, however. The fitting result in testing data can be partially observed in Figure 12. Looking at Figure 10, as larger the observed ozone gets, the more underestimated the prediction is.

When the lags 1 and 2 are removed, that is, only using the lag 24, the metrics get much worse. In special, R^2 is around 0.49. Figure 11 presents how the linear models performs badly when the first lags are removed.

Elastic-net regression

One important observation is that the interaction

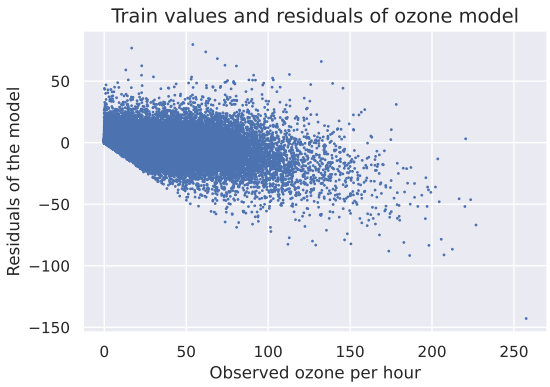


Figure 10: Simple linear regression residuals plotted against observed ozone values.

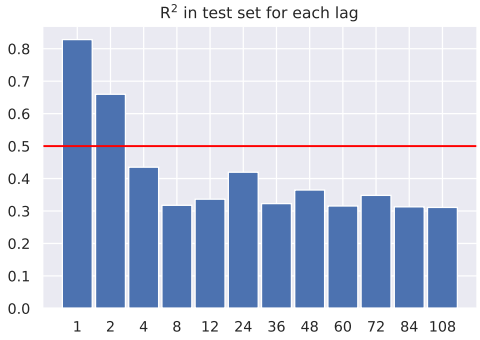


Figure 11: Removing all lag variables, adding only one per time, and calculating the R^2 in the test set.

terms make the metrics worse in the pure linear regression and are disregarded for that reason. In this case, they improve a lot. There are 376 features considering the polynomial and interactions. The hyperparameters were set to $\alpha = 2$ and $w_{l1} = 0.9$, what force several parameters to be zero. Only 6% of the variables ended to be non null. The result was interesting because all the non zero coefficients were related to interactions with the `year` variable. This is related to Figure 3. R^2 in the test set was 0.838, a few more than the simple linear regression. The histogram of the residuals are similar to Figure 9, and skewness and kurtosis too.

Remark. *Without interaction terms, the R^2 was around 0.5 and α was almost 0.*

Feature selection + Linear regression

Feature selection was performed (proper code) with 5-Fold cross validation and threshold equals to 0.001, that is, if the R^2 score improves less than 0.001, the features stop to be added in the best subset. The chosen features were `year * O3_lag1`, `RS`, `RS * hour_cos`, `year * Vel_Vento`, `O3_lag2`, `year O3_MA24`, `hour_cos PM10_lag1`, `RS * Temp`, `RS * O3_lag2`, and `O3_lag1**2`. Note that it is very different from those selected by elastic net. All the variables had p-value less than 0.001, histogram similar to 9, and Jarque-Bera indicating non-normality. The R^2 in testing data was 0.848.

Other important consideration is that the condition number is still big: 9.72e+03, despite being smaller than in simple regression case. After calculating the Pearson correlation between the best variables, `O3_lag2` and `year * O3_lag1` have more than 0.8, what was expected and even used. Some experiments were conducted to reduce this value, but none was successful. It appears that all variables are strongly (and linearly) related but also add some information to prediction (at least a subset of them). This is bad for the interpretation of the coefficients, since its identifiability cannot be proved.

Remark. *Solar radiation had a positive coefficient, so when it increases one unit of standard deviation, the ozone grows too.*

Support Vector Regression

We remember that this model is computationally very costly, so its optimality is not reached. From the cross validation, the best hyperparameters were $\epsilon = 0.2$ and $C = 1.0$. With these values, we perform the linear SVR and the SVR with kernel RBF only to compare them (it is practically impossible to handle this amount of data with grid search when RBF kernel is used in only one computer).

Setting those parameters, SVR with kernel RBF took 6min 28s (CPU time), while Linear SVR 13.1s (CPU time). With RBF kernel, the R^2 in testing

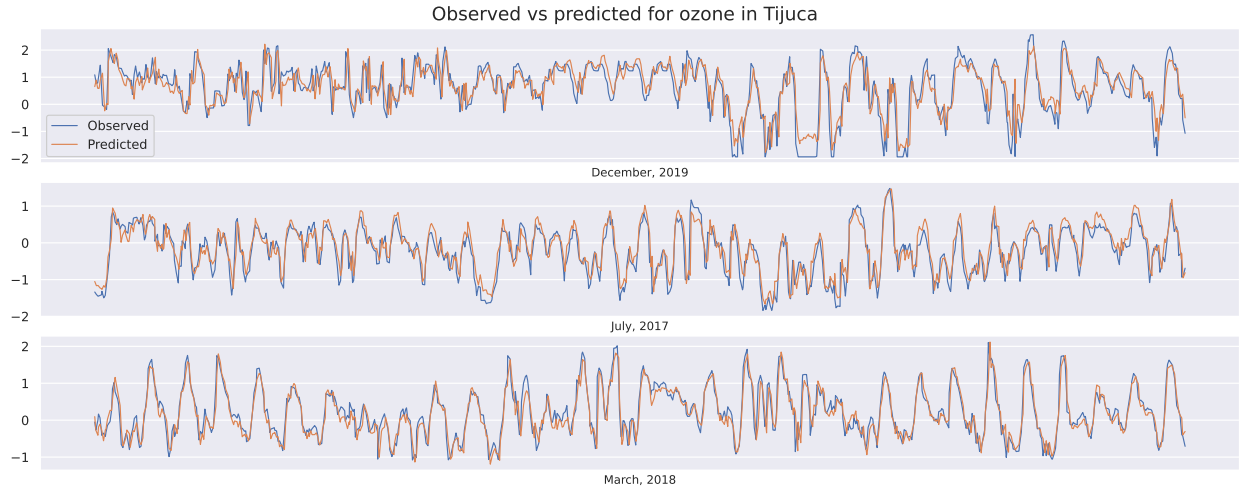


Figure 12: Observed and predicted ozone values for different months in Tijuca.

data was 0.78 and with linear, it was 0.839. Looking at R^2 in training data, we observe that the non-linear kernel suffered with over-fitting. Residues have not changed much from the latest graphs, neither kurtosis nor skewness.

Considering the dataset with polynomial and interaction terms, we used the best features chosen by Forward Feature Selection in the linear regression case. The R^2 in testing data grew to 0.848 (the same as in the linear regression case). The hyperparameters were not changed because a grid search did not change. Feature selection could be done with this model too, but it would be too costly. Finally, the SVR with RBF kernel was performed with the best features. The R^2 in testing set was 0.851.

Random Forest

This algorithm is also very costly for large datasets, specially caused by the choice of s or other parameters, such as, max depth. This makes it harder to do grid search to choose the best parameters. First, we consider that $s \in \{10, 20, 50, 100\}$, $c \in \{0.0, 0.0001, 0.001, 0.01, 0.1, 1\}$, and $B = 100$. The chosen parameters were $s = 10$ and $c = 0$ (no regularization). With this setting, R^2 in testing set was 0.848. From this result, we change the the grid search to $s \in \{2, 5, 8\}$ and $c \in \{0.0, 0.01, 1\}$. As expected, the selected were $c = 0.0$ and $s = 2$. However, despite needing more time, the R^2 in the testing set did not improve.

Linear regression with missing data

The problem with this approach is the absence of a predictive frame. Despite estimating the parameters, the prediction is not straightforward since the data has missing values. To handle this problem, the data with imputation (as explained in Section 5.2) served to predict values. The first thing we should note is that the para meters are different from the linear regression with fixed imputation, but not so much. The ratio between the estimated parameters ranged according to Figure 13.

The R^2 in the testing set was 0.84, a little better than in the simple linear regression case. Considering the best features with polynomial and interactions terms, the R^2 was 0.847.

Model comparison

The best model was the SVR with RBF kernel and the best features (selected by linear model) regarding the three metrics. Figure 14 presents the results for RMSE. Observe the range in y axis goes from 0.2 to 0.36 to emphasize the difference.

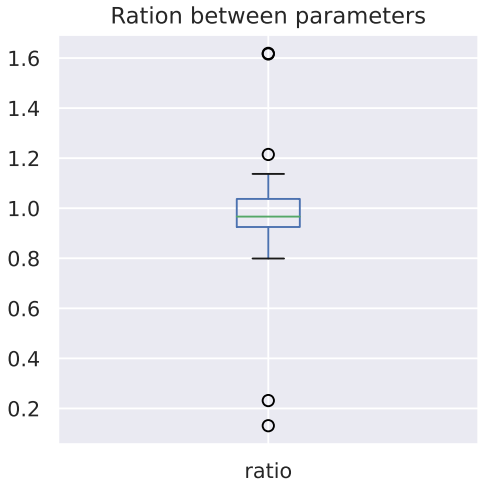


Figure 13: Boxplot with the ratios of the parameters from linear regression with imputation for the parameters with no imputation.

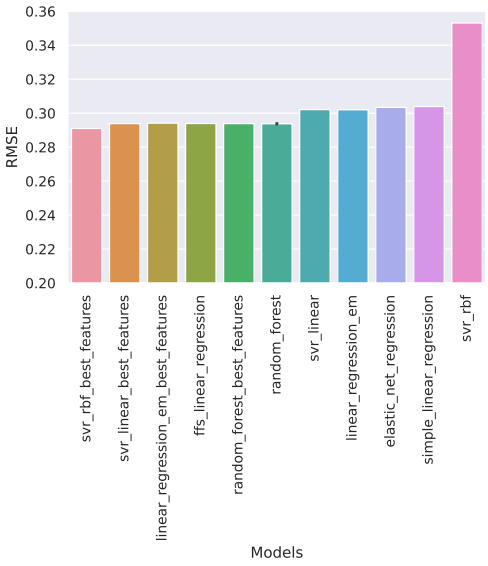


Figure 14: Comparison among the models for ozone concentration at Tijuca monitoring station.

7.1.2 CO

7.1.3 PM_{10}

7.1.4 AIQ

7.2 Aggregated results

7.2.1 O_3

7.2.2 PM_{10}

7.2.3 CO

7.2.4 AIQ

7.3 Model for other locations

Here, we want to make predictions about pollutant levels at other not measured sites. Given that each location has a specific model, the prediction is the weighted mean regarding each prediction

1. Testar estacionaridade de cada série;

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] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

[4] 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.

[5] Harris, C. R., Millman, K. J., and van der Walt*et al*, S. J. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

[6] 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.

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

[8] learn: Machine Learning in Python, S. (2021). sklearn.svm.svr. Available at https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html#sklearn.svm.SVR.

[9] Liang, Y.-C., Maimury, Y., Chen, A. H.-L., and Juarez, J. R. C. (2020). Machine learning-based prediction of air quality. *Applied Sciences*, 10(24):9151.

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

[11] 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.

[12] 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.

[13] pandas development team, T. (2020). pandas-dev/pandas: Pandas.

[14] Pavlo Mozharovskyi, Wei Jiang, M. P. (2021). How to perform parameters estimation with missing values? Available at https://rmisstastic.netlify.app/how-to/estimate/missestim.

[15] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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

[17] 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.

[18] Seabold, S. and Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

[19] 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.

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

[21] Wes McKinney (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

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

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