

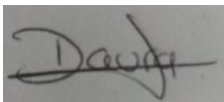
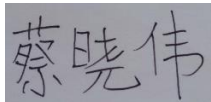
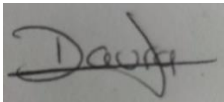
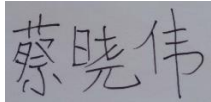
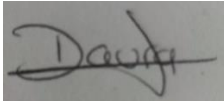
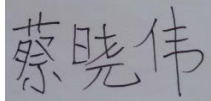
PR2 GENERAL REPORT : CLEANING AND VALIDATION OF DATA

Authors ordered by alphabet:

Daura Hernández Díaz

Xiaowei Cai

URL of GitHub: <https://github.com/likyskyhuuoc/PR2>

Contributions	Signatures	
Previous research		
Writing of the answers		
Code development		

Contents

1. About this study.....	1
2. Introduction to the context.....	1
3. Integration and selection of the data	3
4. Data cleaning.....	5
5. Descriptive data analysis.....	5
6. Inferential data analysis.....	7
1. Statistical assumptions.....	7
2. Correlation analysis.....	8
3. Kruskal–Wallis one-way analysis of variance	9
4. Lineal regression(OLS)	12
5. Generalized additive model(GAM)	13
6. Regression tree	14
7. Predictive model evaluation and selection	15
Model1(Dependent variable: SO2).....	15
Model2(Dependent variable: NO).....	16
Model3(Dependent variable: NO2).....	17
Model4(Dependent variable: NOX)	18
Model5(Dependent variable: O3).....	19
Model6(Dependent variable: CO).....	20
Model7(Dependent variable: PM10).....	21
Model8(Dependent variable: PM2.5).....	22
7. Conclusions	23
8. Code	24
9. Limitations.....	24
Reference	25

1. About this study

This study is an assignment of the course "Typology and life cycle of the data", belonging to Master of Data Science of the Open University of Catalonia. Our group is formed by two persons: Daura Hernández Díaz and Xiaowei Cai, who complicated this study collectively. The report of this study is divided into two parts. The document that you are reading is a general report, which focuses on the written explanation of the research results. On the other hand, we also prepared an implementation report on GitHub¹, where there are the exhaustive coding process and outputs from the R programming environment. We made this distinction in the report in order to allow our readers to find what they want more efficiently.

2. Introduction to the context

Outdoor air pollution is a major environmental health problem which affects everyone in low, middle, and high income countries.

In 2016, environmental (outdoor) air pollution in urban and rural areas is estimated to cause premature deaths of 4.2 million people worldwide due to exposure to particulate matter (PM_{2.5}) of 2.5 microns or less in diameter, which can lead to cardiovascular disease, respiratory diseases, and cancer (WTO, 2018).

People living in low-income and middle-income countries are burdened with increased outdoor air pollution (91% of the 4.2 million premature deaths) occur in these countries, WHO South-East Asia and the Western Pacific. The region faces the heaviest burden. The latest disease burden estimates show that air pollution can have a major impact on cardiovascular disease and death. There is growing evidence that there is a link between environmental air pollution and cardiovascular disease risk, including research from highly polluted areas(WTO, 2018).

WTO (2018) estimates that in 2016, some 58% of outdoor air pollution-related premature deaths were due to ischaemic heart disease and strokes, while 18% of deaths were due to chronic obstructive pulmonary disease and acute lower respiratory infections respectively, and

¹ URL: <https://github.com/likyskyhuuoc/PR2/blob/master/PR2%20Implementation%20Report%20Over%201.1.pdf>

6% of deaths were due to the lung cancer.

Even in a developed country like Spain, air pollution is a serious threat to people's health. More than a decades ago, Spanish researchers began to pay a attention to the air pollution problem, and they note that air pollution rate is significantly associated with several health problems, including asthma (Norris et al., 1999) and cardiovascular diseases(Ballester, Tenías, & Pérez-Hoyos, 2001). Unfortunately, until recent years, air pollution is still a serious health threat in this European country, even in the Canary Islands, which is recognised as the one of the cleanest parts in Spain. Previous studies showed that, in the Canary Islands, there is an association between hospitalizations due to heart failure and exposure to particles in the ambient air (Domínguez-Rodríguez et al., 2011) and SO₂ pollution (Baldasano, Soret, Guevara, Martínez, & Gassó, 2014; Milford, Marrero, Martin, Bustos, & Querol, 2010).

In this study, we pretend to establish several econometric model to predict the pollution level in the Island of Fuerteventura, one of the seven islands in the Canary Islands. The dataset we used in this study is from a governmental website of the Canary Islands². We selected this data source because it is the public database from the government of the Canary Islands, which provides a variety of air quality indicators through a series of monitor stations across the archipelago.

Moreover, in this research we only focus on the data from a specific monitor station located in Fuerteventura (Casa Palacio-Puerto del Rosario). This monitor station was selected due to two reasons. First, this monitor station has the most comprehensive meteorological and pollution indicators we need for our research across a relatively long time. This weather station has been recording data since 2009, although some indicators were not fully recorded at the beginning. Secondly, the purpose of this study is to establish several predictive models instead of explanatory models. In this case, if we use the data across the seven islands in the archipelago, some exogenous variables may be introduced to the models because of the difference among islands.

² URL: <http://www.gobiernodecanarias.org/medioambiente/calidaddelaire/datosHistoricos.do>

Additionally, it is worth mentioning that the crawler which we had created in the PR1 was once again used to simplify the process of obtaining data. When using the crawler, we set up to acquire the daily data from January 2011 to December 2018 in the monitor station of Casa Palacio-Puerto del Rosario with all the meteorological and pollution indicators.

3. Integration and selection of the data

According to the mechanism of the governmental website of the Canary Islands, the maximum time span we can download data is 365 days. Therefore, we need to run our crawler multiple times to scrap the data from year 2011 to 2018. After each execution of the crawler, we acquired four original datasets with different indicators, which contain the data within one year. Finally, 32 original datasets were acquired.

The next step is to integrate these original dataset into one unique integrated dataset. To achieve this goal, we applied R programming language.

In terms of the implementation process, firstly, we use `cbind()` function to integrate the four original datasets of one year to a unique dataset, and we created a new variable named “YEAR” for that later we could track the year of the data. In this phase, we acquired eight datasets named from `Fuerteventura_2011` to `Fuerteventura_2018`. Secondly, the variable names from the original datasets are relatively long, so we create a function called `clean_data()` to rename these variables because the same task need to be conducted eight times. Thirdly, these eight datasets were integrated into one dataset using `rbind()` function, and we named it as *Canarydataset*.

13 different air indicators can be found in *Canarydataset*, which are shown in Figure 1. Moreover, date and time are registered in the dataset.

For detailed information of coding, please refer to the section “*Load data and clean data*” of the implementation report.

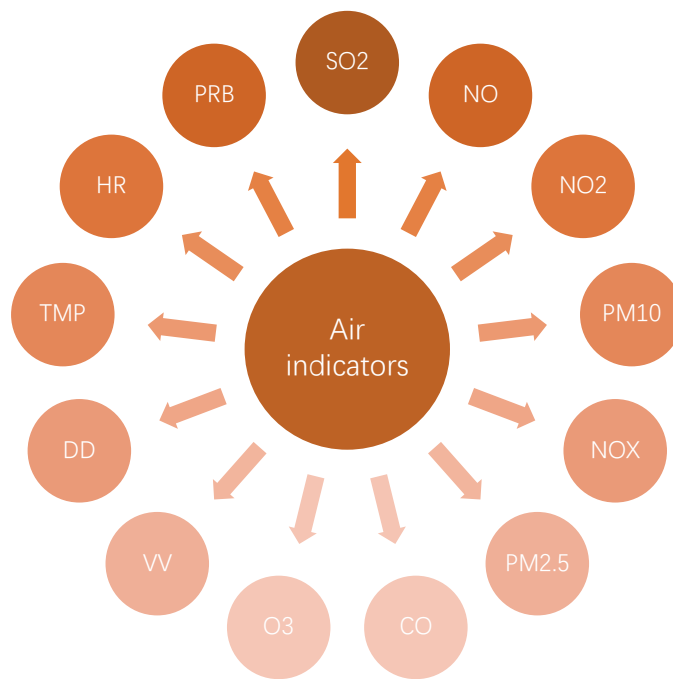


Figure 1 Visualisation of variables in the *Canarydataset*

Here we give the explanations for each variable in the outputs:

DATE: Date when the observation is registered. The format is Day-Month-Year.

HOUR: Time when the observation is registered. 24-hour system.

YEAR: Year when the observation is registered.

MONTH: Month when the observation is registered.

SO2: Concentration of SO2- $\mu\text{g} / \text{m}^3$.

NO: Concentration of NO- $\mu\text{g} / \text{m}^3$.

NO2: NO2- Concentration of NO2- $\mu\text{g} / \text{m}^3$

NOX: Concentration of NOX- $\mu\text{g} / \text{m}^3$.

PM10: Particles in suspension < 10um- $\mu\text{g} / \text{m}^3$.

PM2.5: Particles in suspension < 2,5um- $\mu\text{g} / \text{m}^3$.

CO: Concentration of CO- mg / m^3 .

O3: Concentration of O3- $\mu\text{g} / \text{m}^3$.

VV: Wind speed (Velocidad del viento)- m / s .

DD: Wind direction(Direccion del viento) – Grd.

TMP: Average temperature(Temperatura media)- °C.

HR: Relative humidity(Humedad relativa)-%.

PRB: Barometric pressure(Presion barometrica) – mb.

4. Data cleaning

After having the integrated dataset, we found that in many variables there were NA value. However, we decided to handle this after, because we worried that the intervention of missing values at this moment(e.g. replacing them with the mean value) would be affected by outliers. Moreover, we also detected that there existed some outliers in each of the variables, which may finally bias the further data analysis (Osborne, 2010). We used the argument `$out` in `boxplot.stats()` function to detect the outliers in each of the variables and replaced by NA values. At this moment, the NA values in the dataset come from two sources: the real missing values in the dataset and the NA value we created to represent the outliers.

Then, we replaced all the NA values from both sources with the mean value of each variable. Moreover, we also detected that there were zero values in two variables, PM2.5 and DD. However, these results with a value of 0 are likely to exist in real life, so we will not intervene in these records .

For detailed information of coding, please refer to the section *“Load data and clean data”* of the implementation report.

5. Descriptive data analysis

We calculated a series of descriptive statistics for each variable in the dataset, including mean standard deviation, skewness, kurtosis, and quartiles. Table 1 summarise the descriptive statistics.

Table 1 Descriptive statistics

	mean	sd	skewness	kurtosis	0%	25%	50%	75%	100%	n
SO2	3.161	2.323	1.204	1.032	1.000	1.000	3.000	4.000	11.000	2922
NO	4.370	1.908	0.435	0.311	1.000	3.000	4.370	5.000	10.000	2922
NO2	10.734	4.239	0.561	0.347	1.000	8.000	10.734	13.000	23.000	2922
NOX	17.193	6.216	0.414	0.340	3.000	13.000	17.193	20.000	35.000	2922
O3	72.112	14.780	-0.118	-0.258	31.000	62.000	72.112	83.000	114.000	2922
CO	0.160	0.064	0.791	-0.270	0.100	0.100	0.160	0.200	0.300	2922
PM10	24.469	10.370	0.704	0.312	2.000	17.000	24.000	30.000	57.000	2922
PM2.5	6.347	3.686	0.989	0.633	0.000	3.000	6.000	8.000	18.000	2922
VV	2.847	1.010	-0.117	-0.122	0.100	2.125	2.900	3.500	5.800	2922
DD	95.613	121.065	1.160	-0.373	0.000	21.000	23.000	158.000	337.000	2922
TMP	21.453	2.819	0.275	-0.334	13.800	19.100	21.453	23.700	31.000	2922
HR	64.464	6.338	-0.263	-0.092	47.000	61.000	64.464	69.000	82.000	2922
PRB	1016.131	6.826	0.180	-0.241	999.000	1012.000	1016.000	1020.000	1034.000	2922

Moreover, we generated the histograms for each variables in the dataset to visualisation the statistics mentioned above. These histograms are shown in Figure 2.



Figure 2 Histogram of the variables

6. Inferential data analysis

1. Statistical assumptions

Before conducting any data analysis, we need to check for statistical assumptions. In this study, we firstly check for the normality and homogeneity of the variance, as these two are the very common statistical assumptions when doing parametric modelling.

Firstly, we used the function `mvn()` from MVN package to run the Shapiro-Wilk test for each of the variables in the dataset. Table 2 shows the results of Shapiro-Wilk tests, from which we can see that all the variables are not normally distributed.

Table 2 Results of Shapiro-Wilk test

Test	Continuous variable	Statistic	p value	Normality
Shapiro-Wilk	SO2	0.8458	<0.001	NO
Shapiro-Wilk	NO	0.9556	<0.001	NO
Shapiro-Wilk	NO2	0.969	<0.001	NO
Shapiro-Wilk	NOX	0.9754	<0.001	NO
Shapiro-Wilk	O3	0.9959	<0.001	NO
Shapiro-Wilk	CO	0.7961	<0.001	NO
Shapiro-Wilk	PM10	0.9629	<0.001	NO
Shapiro-Wilk	PM2.5	0.9174	<0.001	NO
Shapiro-Wilk	VV	0.9965	<0.001	NO
Shapiro-Wilk	DD	0.6785	<0.001	NO
Shapiro-Wilk	TMP	0.9842	<0.001	NO
Shapiro-Wilk	HR	0.9907	<0.001	NO
Shapiro-Wilk	PRB	0.9882	<0.001	NO

Moreover, we applied `fligner.test()` function from the car package to conduct Fligner-Killeen Test for each variable across the two categorical variables: YEAR and MONTH. The results are summarised in Table 3, which shows that all the variables are not homogeneous when crossing either categorical variable.

Table 3 Results of Fligner-Killeen

Test	Continuous variable	Categorical variable	Statistic	p value	Homogeneity
Fligner-Killeen	SO2	YEAR	484.83	<0.001	NO
Fligner-Killeen	NO	YEAR	165.62	<0.001	NO
Fligner-Killeen	NO2	YEAR	87.335	<0.001	NO
Fligner-Killeen	NOX	YEAR	70.553	<0.001	NO
Fligner-Killeen	O3	YEAR	58.16	<0.001	NO
Fligner-Killeen	CO	YEAR	88.829	<0.001	NO
Fligner-Killeen	PM10	YEAR	17.873	<0.001	NO
Fligner-Killeen	PM2.5	YEAR	191.29	<0.001	NO
Fligner-Killeen	VV	YEAR	63.402	<0.001	NO
Fligner-Killeen	DD	YEAR	548.27	<0.001	NO
Fligner-Killeen	TMP	YEAR	82.17	<0.001	NO
Fligner-Killeen	HR	YEAR	51.948	<0.001	NO
Fligner-Killeen	PRB	YEAR	196.83	<0.001	NO
Fligner-Killeen	SO2	MONTH	46.776	<0.001	NO
Fligner-Killeen	NO	MONTH	33.019	<0.001	NO
Fligner-Killeen	NO2	MONTH	99.477	<0.001	NO
Fligner-Killeen	NOX	MONTH	36.073	<0.001	NO
Fligner-Killeen	O3	MONTH	26.017	<0.001	NO
Fligner-Killeen	CO	MONTH	69.012	<0.001	NO
Fligner-Killeen	PM10	MONTH	43.01	<0.001	NO
Fligner-Killeen	PM2.5	MONTH	24.026	<0.001	NO
Fligner-Killeen	VV	MONTH	129.82	<0.001	NO
Fligner-Killeen	DD	MONTH	87.189	<0.001	NO
Fligner-Killeen	TMP	MONTH	37.608	<0.001	NO
Fligner-Killeen	HR	MONTH	129.98	<0.001	NO
Fligner-Killeen	PRB	MONTH	65.106	<0.001	NO

Due to that our data does not show the normality nor the homogeneity, we decided to apply non-parametric methods instead of parametric methods when analysing the data.

2. Correlation analysis

Before conducting more complicated statistical operations. We would like to explore the interrelationships among the continuous variables in the dataset. Due to that all the continuous variables in our dataset are not normally distributed, the Spearman's rank correlation

coefficient was used to assess the correlation of the variables. The visualisation of the Spearman's rank correlation coefficient is shown in Figure 3.

From this figure, we can clearly see that some variables are more correlated than others. For example, NO, NO₂, and NO_x are highly correlated, and all these three pollution indicators are negatively associated with VV, which is a meteorological variable. The results of correlation analysis helped us to establish the predictive models in this study, which will be discussed in the following sections.

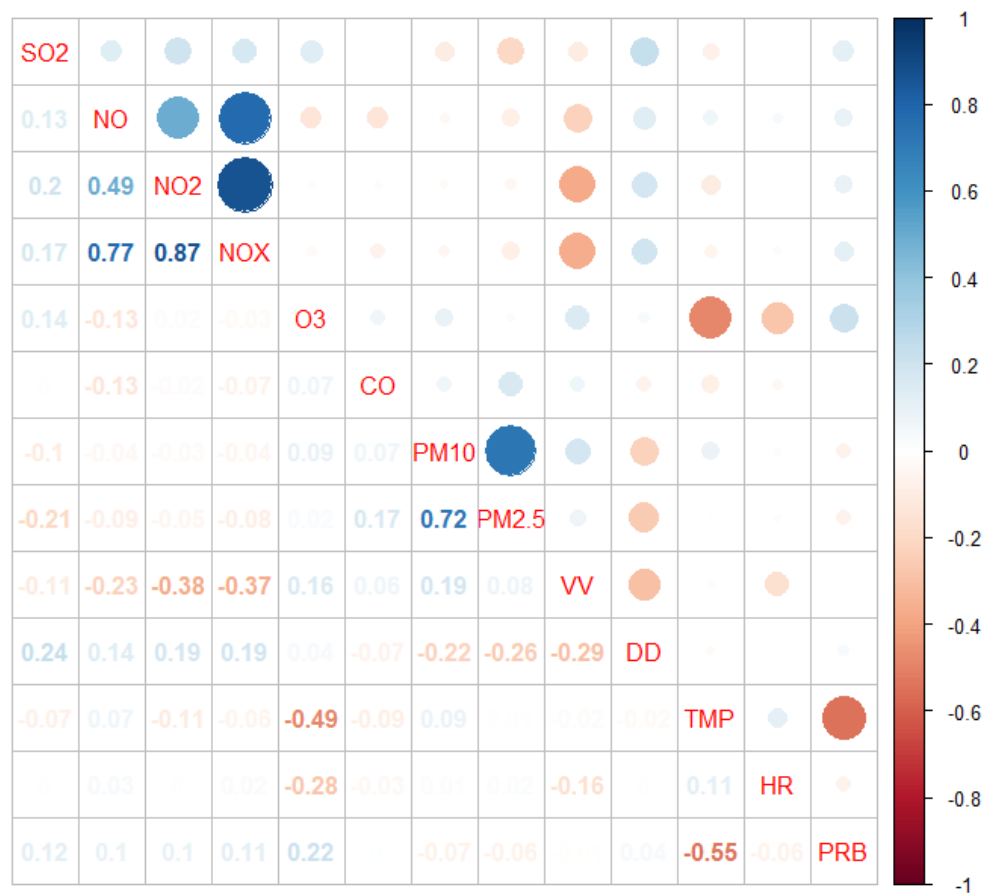


Figure 3 Correlation among variables

3. Kruskal–Wallis one-way analysis of variance

Although the main purpose of this study is to establish econometric models to predict the pollution levels, we also would like to verify whether the pollution indicators are significant different from year 2011 to 2018.

We firstly generated a series of boxplots classified by YEAR and MONTH, which are shown in

Figure 4 and Figure 5. From the visualisation, we can easily infer that the mean value is significant different in many cases. For example, the TMP(temperature) is significantly different across the month. However, statistical tests need to be done to confirm our intuition.

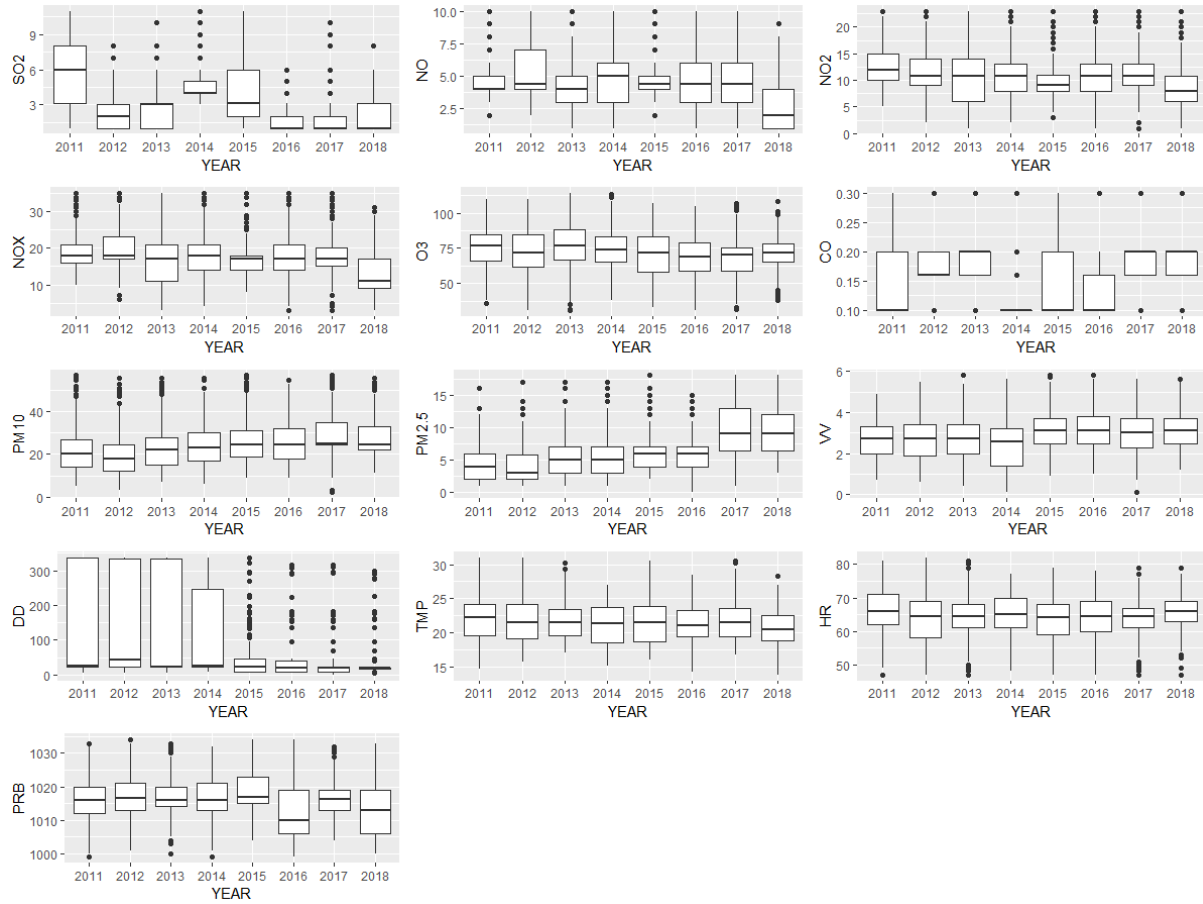


Figure 4 Boxplots by year

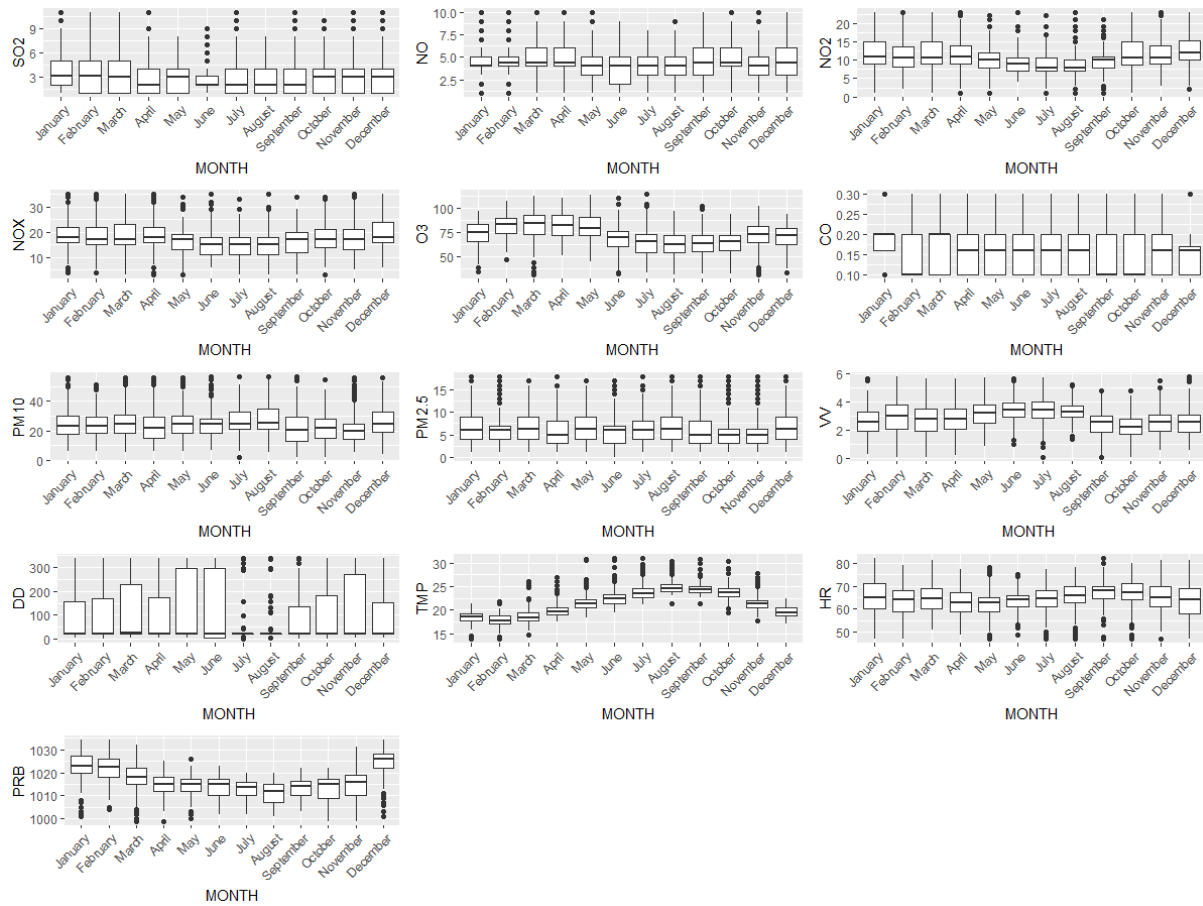


Figure 5 Boxplots by month

In the current study, parametric methods like ANOVA cannot be conducted due to the violation of the statistical assumptions (normality and homogeneity). Thus, a non-parametric method, Kruskal–Wallis test, was adopted to verify whether the mean value is equal in each of the categorical variable. Table 4 shows the results of Kruskal–Wallis test, which indicate that all continuous variables are significantly different respect to the mean value among all the categorical groups.

Table 4 Results of Kruskal-Wallis test

Test	Continues variable	Categorical variable	Statistic	p value
Kruskal-Wallis	SO2	YEAR	1176.6	<0.001
Kruskal-Wallis	NO	YEAR	478	<0.001
Kruskal-Wallis	NO2	YEAR	229.44	<0.001
Kruskal-Wallis	NOX	YEAR	369.87	<0.001

Kruskal-Wallis	O3	YEAR	116.38	<0.001
Kruskal-Wallis	CO	YEAR	631.74	<0.001
Kruskal-Wallis	PM10	YEAR	288.07	<0.001
Kruskal-Wallis	PM2.5	YEAR	875.5	<0.001
Kruskal-Wallis	VV	YEAR	151.66	<0.001
Kruskal-Wallis	DD	YEAR	737.22	<0.001
Kruskal-Wallis	TMP	YEAR	49.284	<0.001
Kruskal-Wallis	HR	YEAR	53.625	<0.001
Kruskal-Wallis	PRB	YEAR	195.25	<0.001
Kruskal-Wallis	SO2	MONTH	44.339	<0.001
Kruskal-Wallis	NO	MONTH	73.909	<0.001
Kruskal-Wallis	NO2	MONTH	302.1	<0.001
Kruskal-Wallis	NOX	MONTH	210.69	<0.001
Kruskal-Wallis	O3	MONTH	765.75	<0.001
Kruskal-Wallis	CO	MONTH	100.35	<0.001
Kruskal-Wallis	PM10	MONTH	107.31	<0.001
Kruskal-Wallis	PM2.5	MONTH	77.617	<0.001
Kruskal-Wallis	VV	MONTH	476.66	<0.001
Kruskal-Wallis	DD	MONTH	75.879	<0.001
Kruskal-Wallis	TMP	MONTH	2244.7	<0.001
Kruskal-Wallis	HR	MONTH	171.35	<0.001
Kruskal-Wallis	PRB	MONTH	1120.2	<0.001

4. Lineal regression(OLS)

Although, as we have mentioned before, our data does not demonstrate the normality and homogeneity, which violate the statistical assumptions of many parametric method. However, we firstly try lineal regression as it is the preferred statistical modelling method for numerical variables (Zumel & Mount, 2014). Researchers should always try linear regression first, and only use more complex methods when more complex methods are better than linear regression models (Zumel & Mount, 2014).

We conducted the log transformation for the variables hoping that the statistical assumptions of least squares method can be satisfied. In this phase, eight regression model were established, and the relationship between dependent and independent variables is based on the previous correlation analysis. These models are shown as following:

Model1: $\ln(\log(\text{SO}_2+1) \sim \text{NO}+\text{NO}_2+\text{NOX}+\text{O}_3+\text{PM}_{2.5}+\text{VV}, \text{data}=\text{Canarydataset_training})$

Model2: $m(\log(\text{NO}+1) \sim \text{NO}_2+\text{NOX}+\text{O}_3+\text{HR}+\text{PRB}, \text{data}=\text{Canarydataset_training})$

Model3: $m(\log(\text{NO}_2+1) \sim \text{SO}_2+\text{NO}+\text{NOX}+\text{O}_3+\text{HR}+\text{PRB}, \text{data}=\text{Canarydataset_training})$

Model4: $m(\log(\text{NOX}+1) \sim \text{SO}_2+\text{NO}+\text{NO}_2+\text{O}_3+\text{HR}+\text{PRB}, \text{data}=\text{Canarydataset_training})$

Model5: $m(\log(\text{O}_3+1) \sim \text{SO}_2+\text{NO}+\text{NO}_2+\text{NOX}+\text{HR}+\text{PRB}, \text{data}=\text{Canarydataset_training})$

Model6: $m(\log(\text{CO}+1) \sim \text{NO}+\text{NO}_2+\text{NOX}+\text{PM}_{2.5}+\text{DD}+\text{HR}, \text{data}=\text{Canarydataset_training})$

Model7: $m(\log(\text{PM}_{10}+1) \sim \text{NO}+\text{PM}_{2.5}+\text{DD}+\text{TMP}+\text{HR}+\text{PRB}, \text{data}=\text{Canarydataset_training})$

Model8: $m(\log(\text{PM}_{2.5}+1) \sim \text{SO}_2+\text{PM}_{10}+\text{VV}+\text{DD}+\text{TMP}, \text{data}=\text{Canarydataset_training})$

Moreover, we used `gvlma()` function from the `gvlma` package (Peña & Slate, 2006) to verify the statistical assumptions of least squares method. However, unfortunately, according to the method of Peña & Slate (2006), all these eight lineal regression models do not satisfy the statistical assumptions, which obliged us to seek for the more complex non-parametric methods to achieve our research goal.

For detailed information of coding and estimation results, please refer to the section “*Linal regression models*” of the implementation report.

5. Generalized additive model(GAM)

As our dataset violate several statistical assumptions that prevented us from using parametric methods, we turned to some non-parametric statistical method and machine learning algorithm to achieve our goal.

An important statistical development of the last 30 years has been the advance in regression analysis provided by generalized additive models (GAMs) (Guisan, Edwards, Jr, & Hastie, 2002). GAMs are semi-parametric extensions of GLMs, and the only underlying assumption made is that the functions are additive and that the components are smooth. The strength of GAMs is their ability to deal with highly non-linear and non-monotonic relationships between the response and the set of explanatory variables (Guisan et al., 2002).

In the R environment, GAMs can be implemented by using `gam()` function of `mgcv` package. Before running the GAM, we randomly separated the dataset into training set(*Canarydataset_training*) and verification set(*Canarydataset_verification*), and in each subset, there are 50% of the total observations. Additionally, `set.seed()` function was used so

that we could replicate the results in the future. These two subsets will also be used in the regression tree analysis, which will be explained in the next section.

The models we specified for GAMs are the same as we did before for lineal regressions, which are shown as following:

Model1: `s(NO)+s(NO2)+s(NOX)+s(O3)+s(PM2.5)+s(VV),data=Canarydataset_training)`

Model2: `gam(NO ~ s(NO2)+s(NOX)+s(O3)+s(HR)+s(PRB),data=Canarydataset_training)`

Model3: `gam(NO2 ~ (SO2)+s(NO)+s(NOX)+s(O3)+s(HR)+s(PRB),data=Canarydataset_training)`

Model4: `gam(NOX ~ s(SO2)+s(NO)+s(NO2)+s(O3)+s(HR)+s(PRB),data=Canarydataset_training)`

Model5: `gam(O3 ~ s(SO2)+s(NO)+s(NO2)+s(NOX)+s(HR)+s(PRB),data=Canarydataset_training)`

Model6: `gam(CO ~`

`s(NO)+s(NO2)+s(NOX)+s(PM2.5)+s(DD)+s(HR),data=Canarydataset_training)`

Model7: `gam(PM10 ~`

`(NO)+s(PM2.5)+s(DD)+s(TMP)+s(HR)+s(PRB),data=Canarydataset_training)`

Model8: `gam(PM2.5 ~ s(SO2)+s(PM10)+s(VV)+s(DD)+s(TMP),data=Canarydataset_training)`

Comprehensive results of these GAMs are not going to be displayed here due to space reasons.

For detailed information of coding and estimation results, please refer to the section “*Generalized additive model(GAM)*” of the implementation report.

6. Regression tree

GAM is a powerful statistical tool to analysis our dataset, but it is not the only method. However, if we shift our focus from statistics to machine learning, there may be more options to solve the same problem. Among all the possible machine leaning algorithm, maybe regression tree is more suitable for our study.

The regression tree was introduced as part of the classification regression tree algorithm(CART) in the 1980s (Breiman, Friedman, Olshen, & Stone, 1984). The regression tree does not use a linear regression method, but instead makes predictions based on the average of the cases that arrive at the leaf nodes (Lantz, 2015). In some cases, the fit of some kinds of data is much better than lineal regression model (Lantz, 2015).

In the R environment, regression tree can be implemented by using `rpart()` function in the `rpart`

package.

The models we specified for regression trees are the same as we did before for lineal regressions, which are shown as following:

Model1:`rpart(SO2 ~NO+NO2+NOX+O3+PM2.5+VV,data=Canarydataset_training)`

Model2:`rpart(NO ~NO2+NOX+O3+HR+PRB,data=Canarydataset_training)`

Model3:`rpart(NO2 ~SO2+NO+NOX+O3+HR+PRB,data=Canarydataset_training)`

Model4:`rpart(NOX ~SO2+NO+NO2+O3+HR+PRB,data=Canarydataset_training)`

Model5:`rpart(O3 ~SO2+NO+NO2+NOX+HR+PRB,data=Canarydataset_training)`

Model6:`rpart(CO ~NO+NO2+NOX+PM2.5+DD+HR,data=Canarydataset_training)`

Model7:`rpart(PM10 ~NO+PM2.5+DD+TMP+HR+PRB,data=Canarydataset_training)`

Model8:`rpart(PM2.5 ~ SO2+PM10+VV+DD+TMP,data=Canarydataset_training)`

Comprehensive results of these regression trees are not going to be displayed here due to space reasons. For detailed information of coding and estimation results, please refer to the section “Regression tree” of the implementation report.

7. Predictive model evaluation and selection

Now we have eight GAM models and eight regression tree models, and at this moment we would like to know which of the two types of approaches functions better for each of the dependent variables. We used `predict()` function with our verification set(*Canarydataset_verification*) to evaluate the performance of the established models. In the current study, the evaluation criterion is based on NMSE(Normalised Mean Square Error)³, and scatter plots helps us to understand the results more intuitively.

Model1(Dependent variable: SO2)

In the models where SO2 is a dependent variable, the NMSE value for GAM model is 0.8669353 and the NMSE value for regression tree model is 0.9242515. Thus, neither of the two models

³ The NMSE (Normalised Mean Square Error) is an estimator of the overall deviations between predicted and measured values. NMSE is a ratio and usually ranges from 0 to 1. If the model performs well, the NMSE value should be significantly less than one. That is to say, the smaller the value of NMSE, the better the performance of the model.

performs well, although the performance of the GAM model is slightly better. Figure 6 compares the GAM model and the regression tree model.

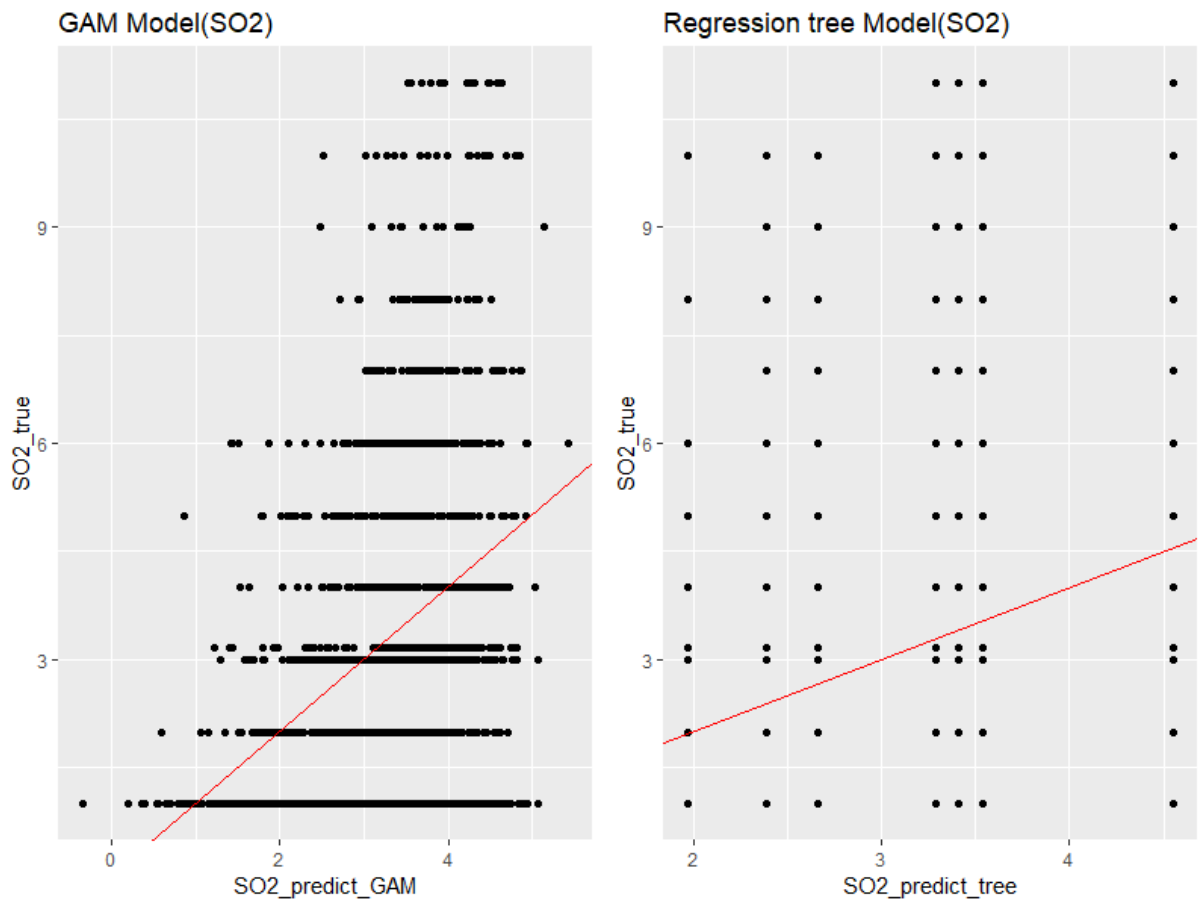


Figure 6 Model evaluation(SO2 as the dependent variable)

Model2(Dependent variable: NO)

In the models where NO is a dependent variable, the NMSE value for GAM model is 0.2903619 and the NMSE value for regression tree model is 0.32765. Thus, both of the two models perform well, although the performance of the GAM model is slightly better. Figure 7 compares the GAM model and the regression tree model.

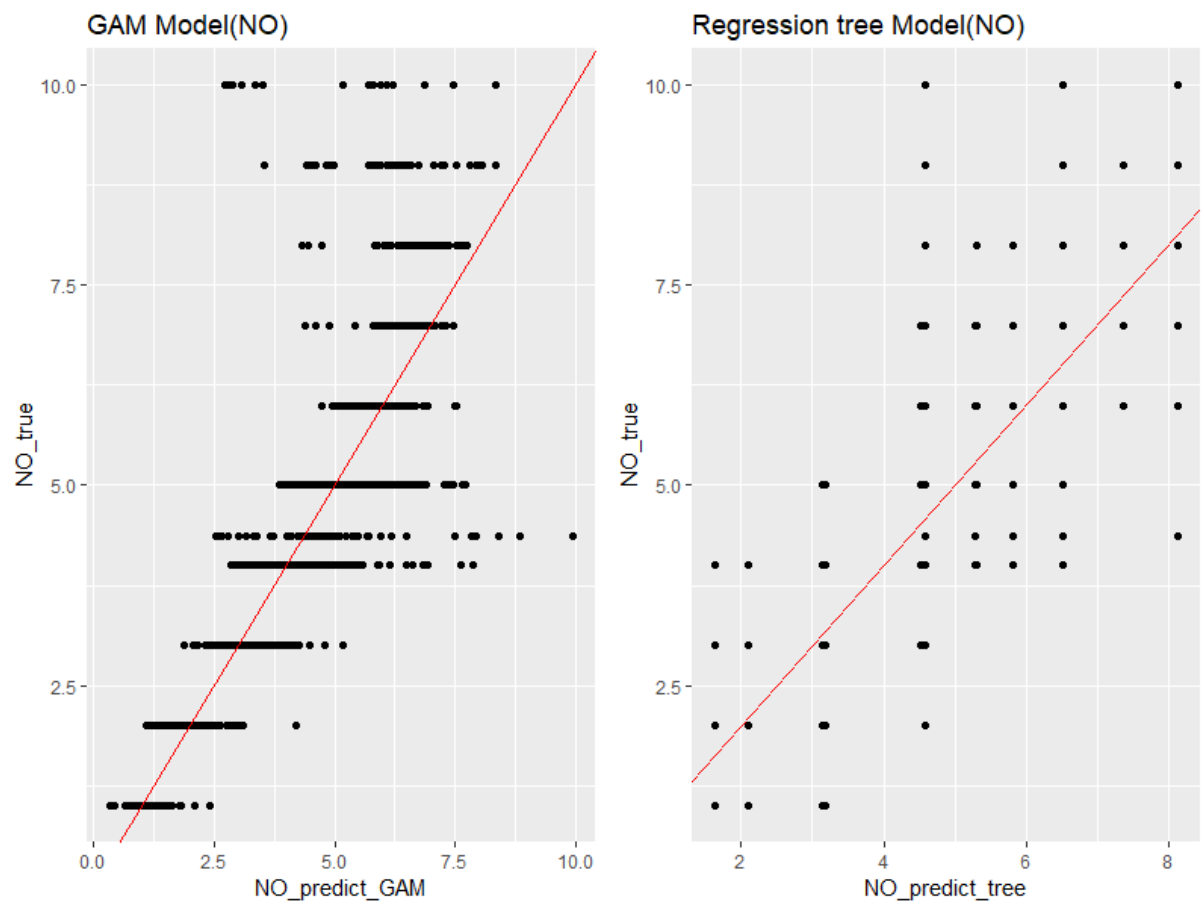


Figure 7 Model evaluation(NO as the dependent variable)

Model3(Dependent variable: NO2)

In the models where NO2 is a dependent variable, the NMSE value for GAM model is 0.1996935 and the NMSE value for regression tree model is 0.270828. Thus, both of the two models perform well, although the performance of the GAM model is slightly better. Figure 8 compares the GAM model and the regression tree model.

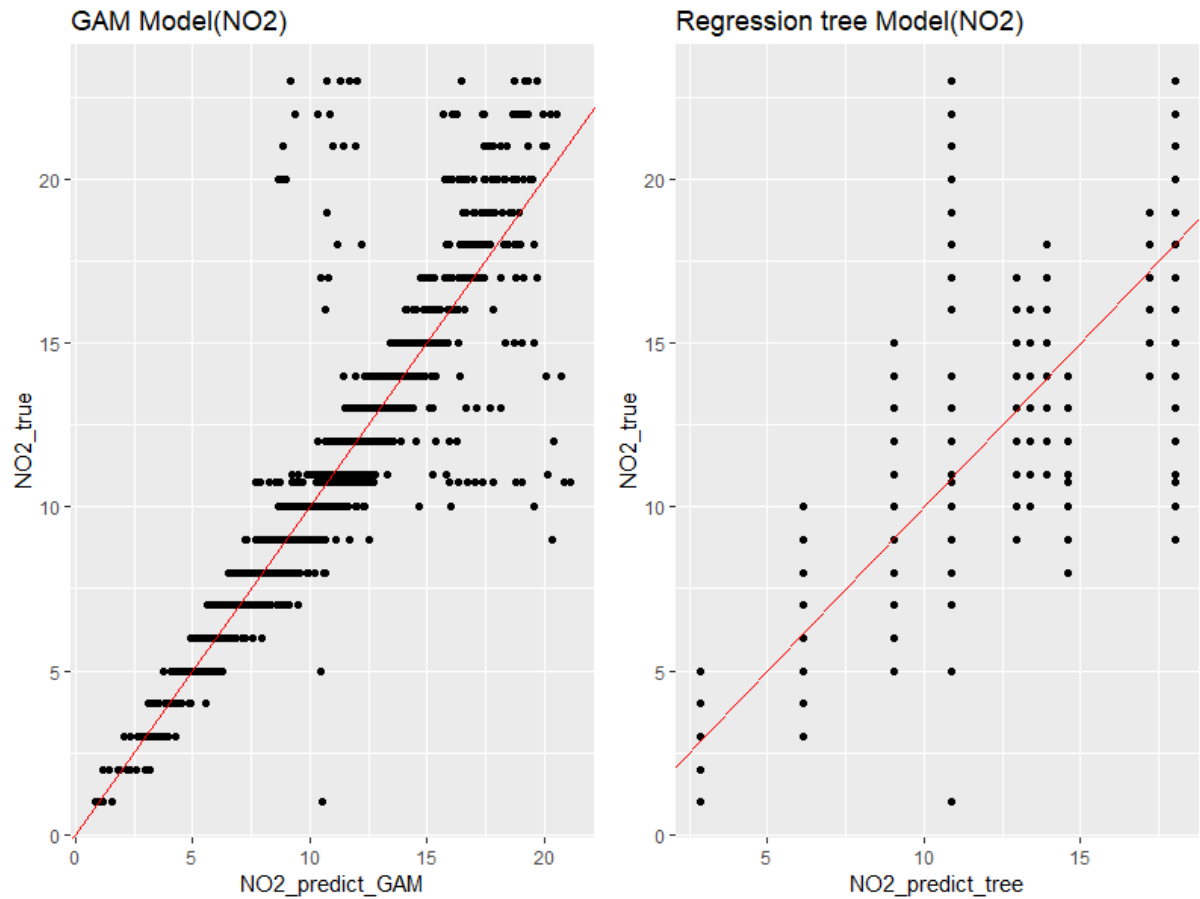


Figure 8 Model evaluation(NO2 as the dependent variable)

Model4(Dependent variable: NOX)

In the models where NOX is a dependent variable, the NMSE value for GAM model is 0.124064 and the NMSE value for regression tree model is 0.2002506. Thus, both of the two models perform well, although the performance of the GAM model is slightly better. Figure 9 compares the GAM model and the regression tree model.

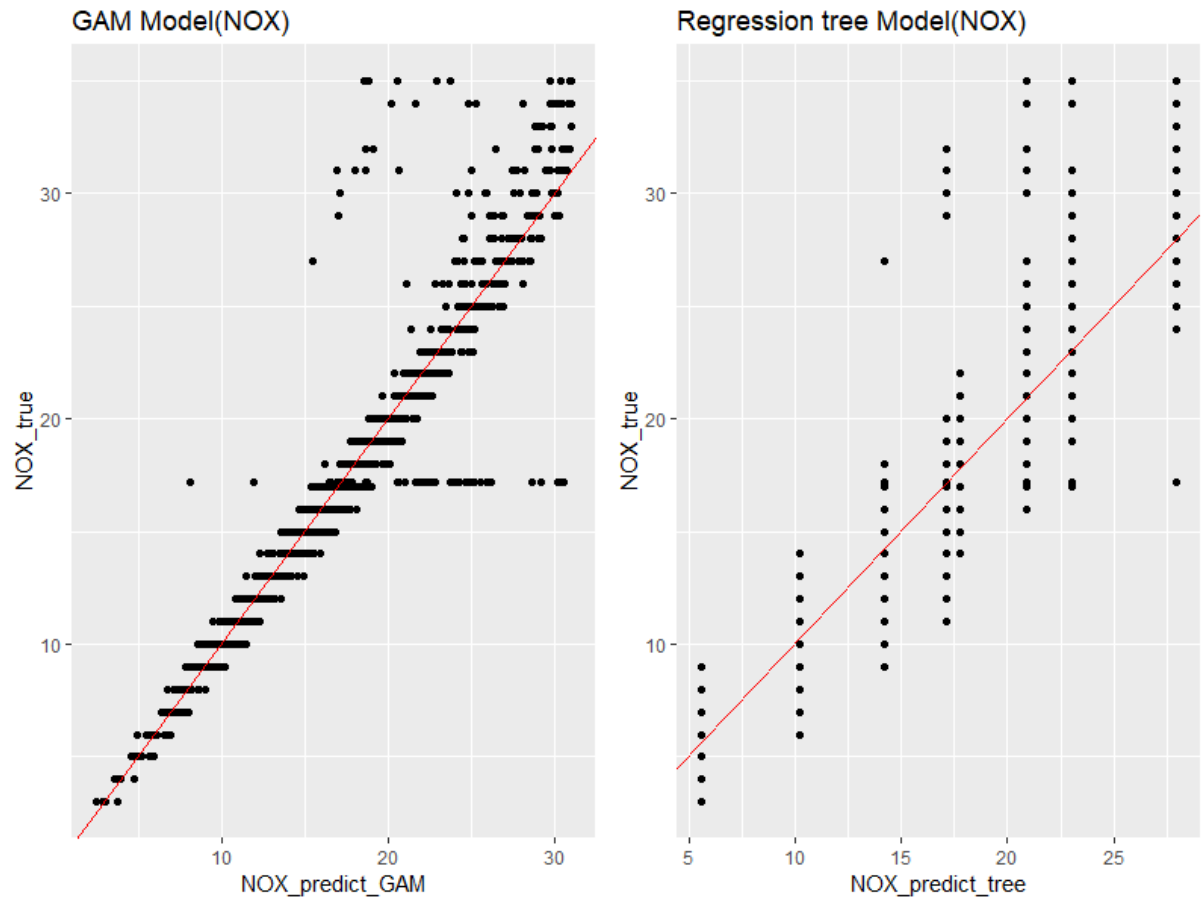


Figure 9 Model evaluation(NOX as the dependent variable)

Model5(Dependent variable: O3)

In the models where O3 is a dependent variable, the NMSE value for GAM model is 0.8522244 and the NMSE value for regression tree model is 0.8485102. Thus, neither of the two models performs well, although the performance of the tree model is slightly better. Figure 10 compares the GAM model and the regression tree model.

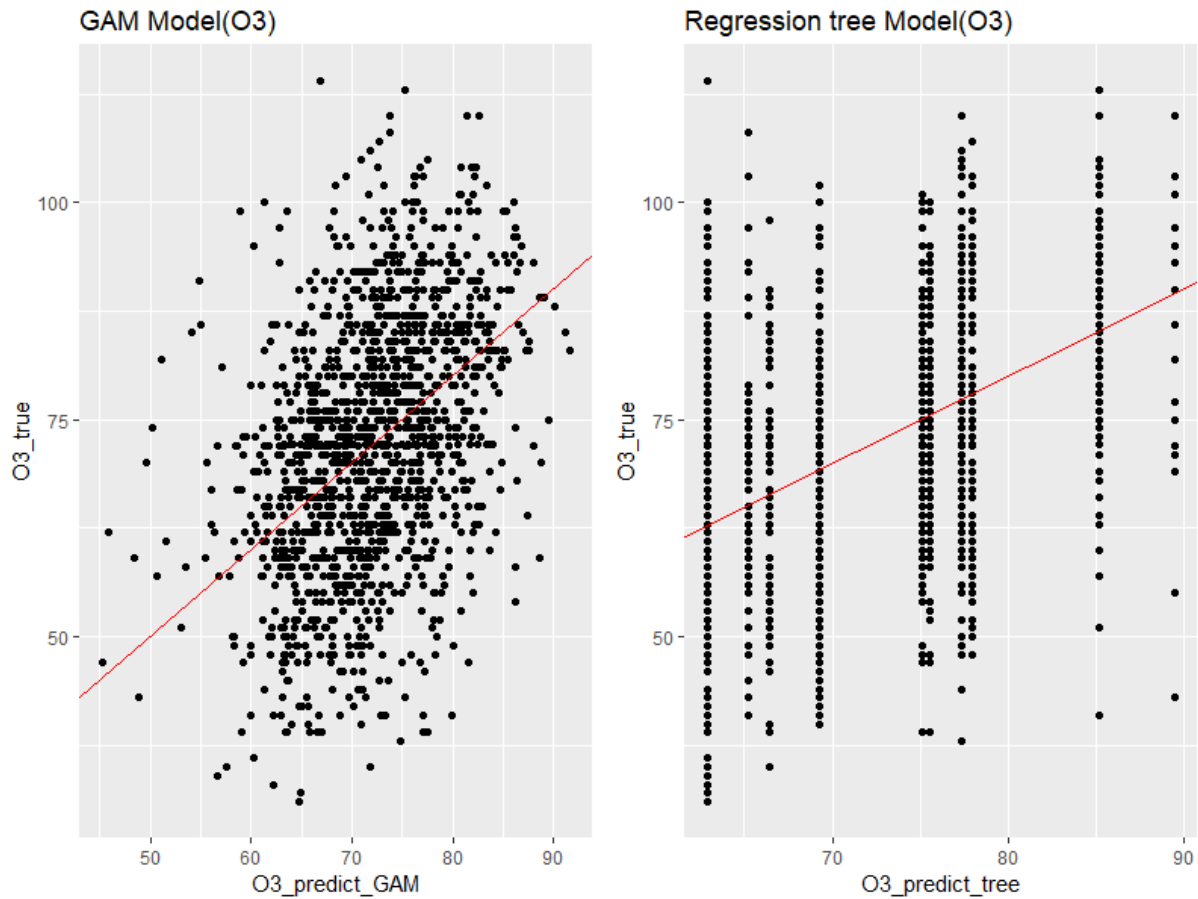


Figure 10 Model evaluation(O3 as the dependent variable)

Model6(Dependent variable: CO)

In the models where CO is a dependent variable, the NMSE value for GAM model is 0.9556282 and the NMSE value for regression tree model is 0.945722. Thus, neither of the two models performs well, although the performance of the tree model is slightly better. Figure 11 compares the GAM model and the regression tree model.

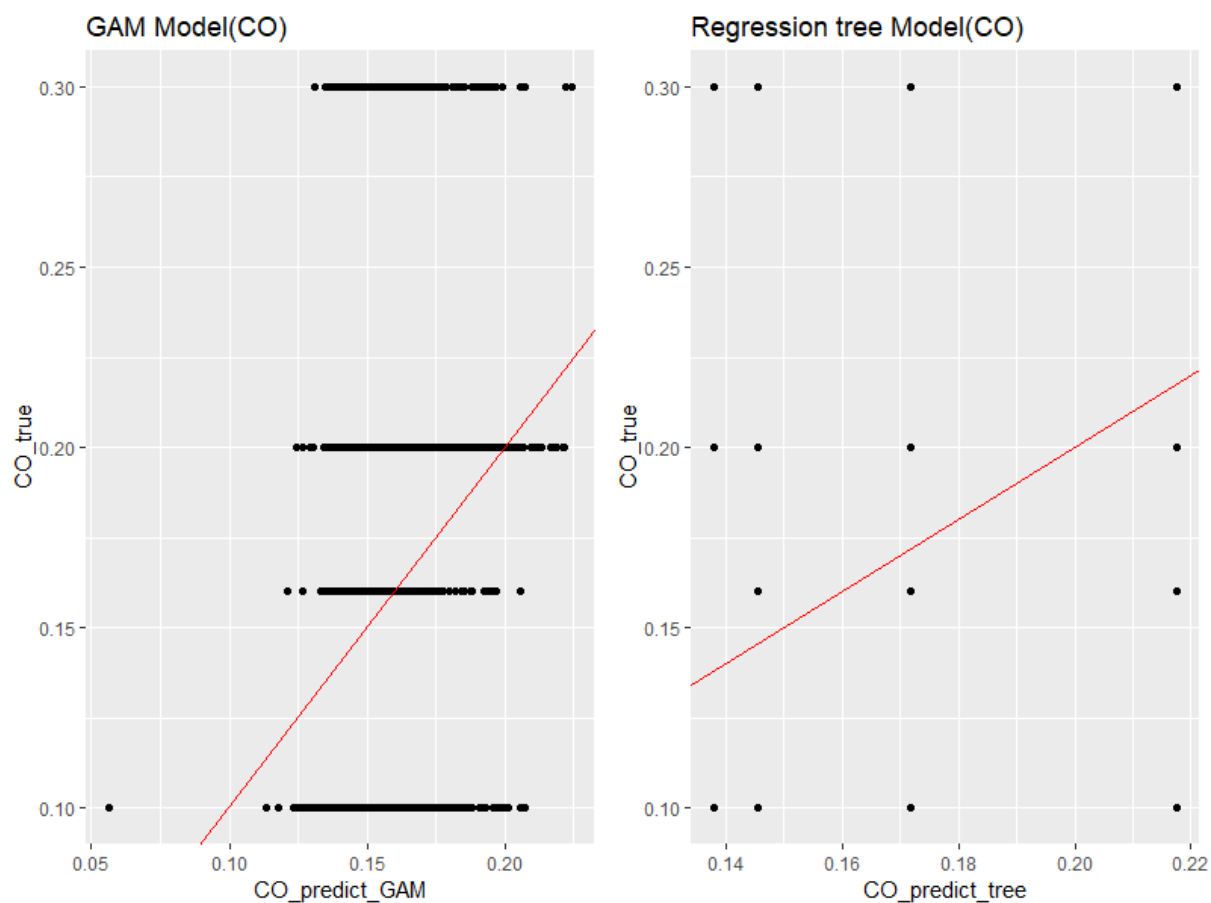


Figure 11 Model evaluation(CO as the dependent variable)

Model7(Dependent variable: PM10)

In the models where PM10 is a dependent variable, the NMSE value for GAM model is 0.497423 and the NMSE value for regression tree model is 0.5187397. Thus, neither of the two models perform generally well, although the performance of the GAM model is slightly better. Figure 12 compares the GAM model and the regression tree model.

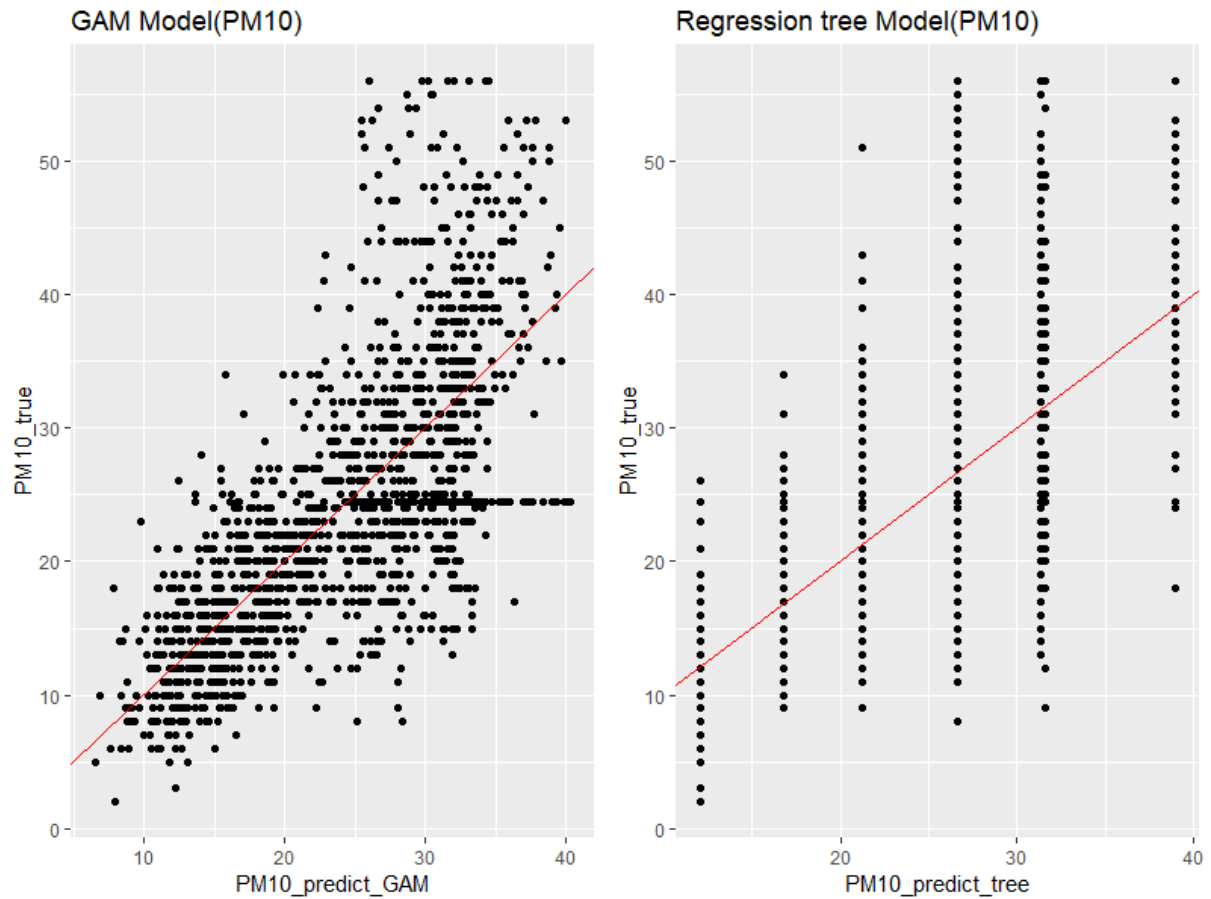


Figure 12 Model evaluation(PM10 as the dependent variable)

Model8(Dependent variable: PM2.5)

In the models where SO₂ is a dependent variable, the NMSE value for GAM model is 0.5230972 and the NMSE value for regression tree model is 0.5680897. Thus, both of the two models perform generally well, although the performance of the GAM model is slightly better. Figure 12 compares the GAM model and the regression tree model.

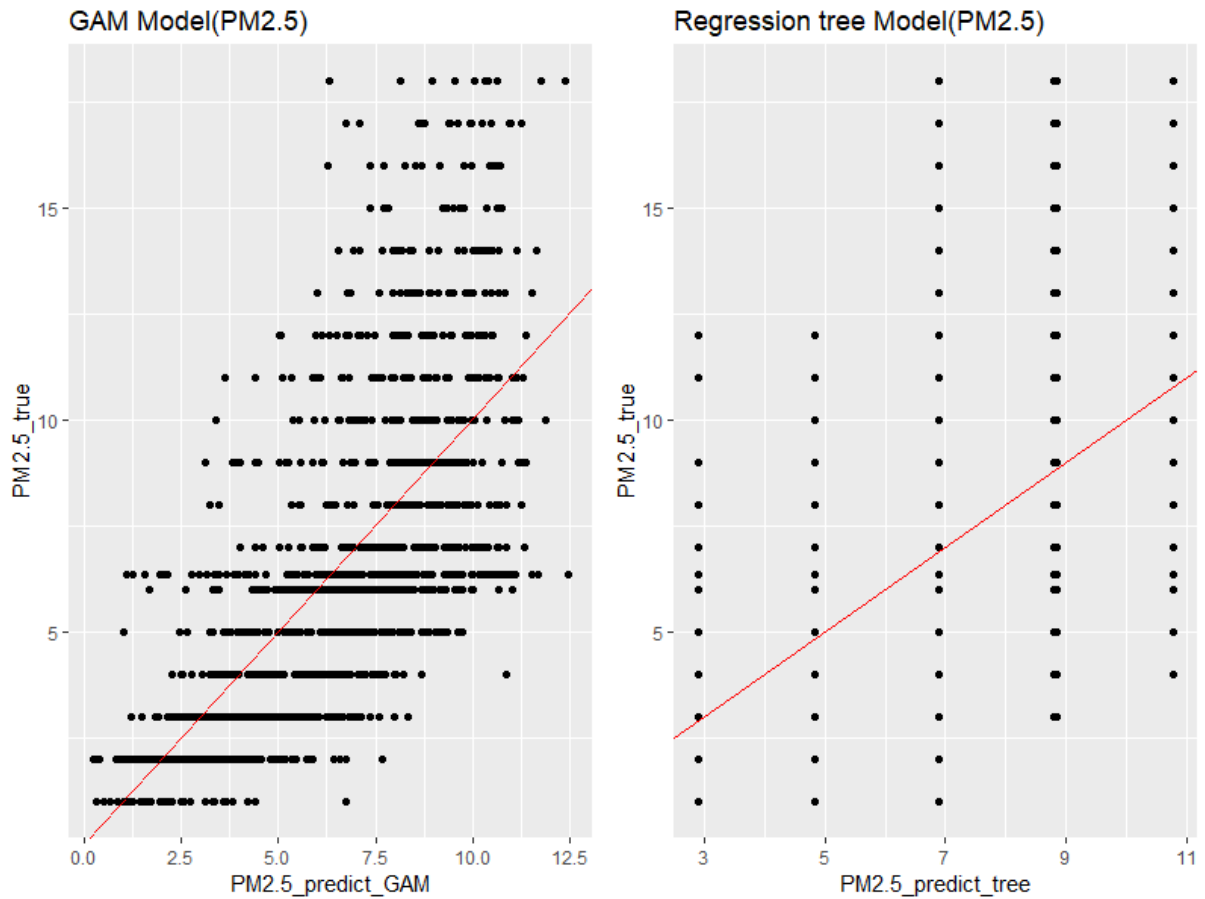


Figure 13 Model evaluation(PM2.5 as the dependent variable)

7. Conclusions

In this research, we aimed to establish predictive models to forecast the pollution level in Fuerteventura, one of the seven islands in the Canary Islands. To reach this goal, we used the crawler that we had created in the PR1 to acquire the datasets we wanted, and later we used R programming language to integrate, clean, and analyse the data. Several statistical and machine learning methods were used in this research. Firstly, we conducted descriptive statistical analysis to acquire some basic information about our dataset. Then, Shapiro-Wilk tests and Fligner-Killeen tests were conducted to test the normality and homogeneity of the data. Moreover, correlation analysis using Spearman's rank correlation coefficient helped us to explore the relationships among the variables. Additionally, Kruskal-Wallis tests were conducted to verify whether the mean value of the meteorological and pollution variables were significantly different across the year and month. In terms of the inferential statistic, we transformed

dependent variables to run lineal regression models with ordinary least squares estimation. However, even the dependent variables were transformed, these lineal regression models did not meet the statistical assumptions. Due to this reason, Generalized additive model and Regression tree were used in this study as two alternative methods to establish predictive models. As the final step, we compare the fits of model using Normalised Mean Square Error value and visualisation. We found that the models to predict NO, NO₂, and NO_x were well fitted. The models to predict PM₁₀ and PM_{2.5} were generally well fitted. Nevertheless, the models to predict SO₂, O₃, and CO were not well fitted. In conclusion, we partially reached our research goals

8. Code

Due to the space and simplicity reasons, the source R code will not be displayed here. In turn, our readers can find the full code as well as the original datasets of this project in GitHub⁴. We highly encourage our readers to check the implementation report to understand the entire implementation process of this project.

9. Limitations

Like all investigations, our research is subject to some limitations. Maybe one of the biggest limitation of this study is that the two authors of this study are not specialist in the field of environmental science. As a result, no literature review has been done before conducting the data analysis, which may lead to some errors that are closely related to subject knowledge. Additionally, we tried to avoid interpreting the results of data analysis because the process of interpretation requires a great need of subject knowledge. Finally, all the datasets we used in this study were from a specific monitor station() in Fuerteventura, which may limit the generalisability to the whole island.

⁴ URL: <https://github.com/likyskyhuuoc/PR2>

Reference

- Baldasano, J. M., Soret, A., Guevara, M., Martínez, F., & Gassó, S. (2014). Integrated assessment of air pollution using observations and modelling in Santa Cruz de Tenerife (Canary Islands). *Science of the Total Environment*.
<https://doi.org/10.1016/j.scitotenv.2013.12.062>
- Ballester, F., Tenías, J. M., & Pérez-Hoyos, S. (2001). Air pollution and emergency hospital admissions for cardiovascular diseases in Valencia, Spain. *Journal of Epidemiology and Community Health*. <https://doi.org/10.1136/jech.55.1.57>
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees* (1st ed.). New York: Routledge. <https://doi.org/10.1201/9781315139470>
- Domínguez-Rodríguez, A., Abreu-Afonso, J., Rodríguez, S., Juárez-Prera, R. A., Arroyo-Ucar, E., Jiménez-Sosa, A., ... Avanzas, P. (2011). Comparative Study of Ambient Air Particles in Patients Hospitalized for Heart Failure and Acute Coronary Syndrome. *Revista Española de Cardiología (English Edition)*. <https://doi.org/10.1016/j.rec.2010.12.023>
- Guisan, A., Edwards, T. C., Jr, & Hastie, T. (2002). Generalized linear and generalized additive models in studies of species *Ecological Modelling*, 157. Retrieved from <http://linkinghub.elsevier.com/retrieve/pii/S0304380002002041>
- Lantz, B. (2015). *Machine Learning with R* (2nd ed.). Packt Publishing.
- Milford, C., Marrero, C., Martin, C., Bustos, J. J., & Querol, X. (2010). Forecasting the air pollution episode potential in the Canary Islands. *Advances in Science and Research*.
<https://doi.org/10.5194/asr-2-21-2008>
- Norris, G., YoungPong, S. N., Koenig, J. Q., Larson, T. V., Sheppard, L., & Stout, J. W. (1999). An association between fine particles and asthma emergency department visits for children in Seattle. *Environmental Health Perspectives*. <https://doi.org/10.1289/ehp.99107489>
- Osborne, J. W. (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. *Newborn and Infant Nursing Reviews*, 10(1), 37–43.
<https://doi.org/10.1053/j.nainr.2009.12.009>

- Peña, E. A., & Slate, E. H. (2006). Global Validation of Linear Model Assumptions. *Journal of the American Statistical Association*, 101(473), 341–354. Retrieved from <https://amstat.tandfonline.com/doi/abs/10.1198/016214505000000637#.XHRiulhKjIU>
- WTO. (2018). Ambient (outdoor) air quality and health. Retrieved from [https://www.who.int/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health](https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health)
- Zumel, N., & Mount, J. (2014). *Practical data science with R* (1st ed.). Manning Publications Co.