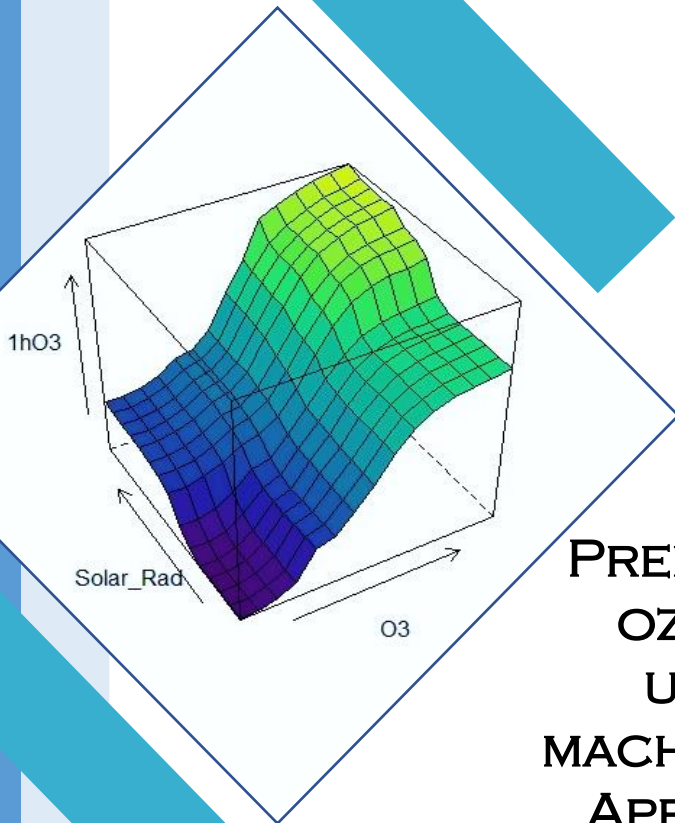


MASTER'S THESIS



PREDICTION OF TROPOSPHERIC OZONE CONCENTRATION AT URBAN LOCATIONS USING MACHINE LEARNING ALGORITHMS. APPLICATION TO BARCELONA, SPAIN

SERGIO RICARDO LÓPEZ CHACÓN

ACADEMIC SUPERVISOR: MANUEL GOMEZ VALENTÍN

INSTITUTIONAL SUPERVISOR : FERNANDO SALAZAR GONZÁLEZ



A mis padres, la luz de mi vida

Acknowledgement

I would like to express my deep gratitude to Fernando Salazar who guided me in every step of this master's thesis. I learned a lot from him and without his contribution and collaboration, this work would not have been possible.

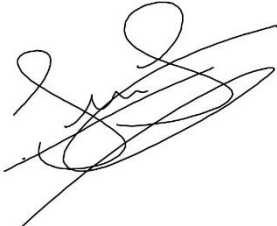
I also thank Prof. Manuel Gomez, Euroaquae coordinator in Barcelona for all the support and advice that he gave me while I had the opportunity to study at the Polytechnic University of Catalonia.

This master's degree has been the greatest experience of my life and I am deeply grateful to Prof. Frank Molkenhain who since the beginning of that has helped me, guided me and supported me.

I also thank my parents who are the engine of my life, in the happiest and saddest moments they are always there for me. Finally, my friends who I consider part of my family too.

Declaration

I declare that this master's thesis was written by myself without any unauthorized support, and it has not been used to obtain a professional title previously. The information obtained from external sources has been properly cited.

A handwritten signature in black ink, consisting of several loops and strokes, positioned above the printed name.

Sergio López Chacón

Barcelona. August 15th, 2021

Table of contents

1	Introduction and objectives	1
1.1	Aim	2
1.2	Overview of the methodology	3
2	Literature Review	4
2.1	Ground-level ozone	4
2.2	Characteristics of the ozone and meteorological stations	4
2.2.1	Ozone measurements	5
2.2.2	Automatic meteorological stations	5
2.3	Machine learning (ML)	5
2.4	State-of-the-art on the use of ML for predicting O ₃ levels	6
2.5	Decision trees and recursive partitioning	8
2.6	Random forest (RF)	9
2.7	Error metrics	10
2.8	Overfitting	11
2.9	Cross validation	11
2.10	Prequential evaluation method	12
2.11	Categorical models and confusion matrix	13
2.12	R libraries for RF	14
3	Study case	16
3.1	Study area	16
3.2	Stations and data available	17
3.2.1	Source	18
3.2.2	Time window for Barcelona-Palau Reial and Barcelona-Zona Universitaria (PR_ZU)	18
3.2.3	Time window for Barcelona-Eixample and Barcelona-El Raval (EI_RA)	19
4	Methodology	21
4.1	Pre-processing	21
4.1.1	Missing values and Imputing data	21
4.1.2	Outputs	22
4.1.3	Inputs	23
4.1.4	Procedure to obtain output and inputs	24
4.2	Exploration	27
4.2.1	Outputs	27

4.2.2	Trends and linear correlations between outputs and inputs	30
4.3	Training	35
4.4	Testing	37
5	Results	39
5.1	Models of the whole year	39
5.1.1	Regression models.....	39
5.1.2	Categorical model	49
5.2	Models from May to September.....	54
5.2.1	Regression models.....	54
5.2.2	Categorical model	63
6	Conclusions.....	68
6.1	Regression models.....	68
6.2	Categorical models	69
	References	71

Table of figures

Figure 1 Atmosphere layers (UCAR Centre for Science Education, 2021).....	1
Figure 2 ML models workflow	6
Figure 3 Scheme of Bootstrap sampling	9
Figure 4 Underfitting, good fit and overfitting (Ghojogh and Crowley, 2019)	11
Figure 5 K-folds cross validation	12
Figure 6 Schemes of Prequential evaluation method	12
Figure 7 Scheme of the confusion matrix	13
Figure 8 Main parameters of <i>randomForest</i> function	14
Figure 9 Main parameters of <i>cforest</i> function	14
Figure 10 Location of the selected stations in Barcelona.....	16
Figure 11 Time series of air quality variables data available of Barcelona-Palau Reial station	18
Figure 12 Time series of meteorological variables data available of Barcelona-Zona Universitaria station	19
Figure 13 Time series of air quality variables data available of Barcelona-Eixample station	20
Figure 14 Time series of meteorological variables data available of Barcelona-El Raval station	20
Figure 15 Steps of the methodology.....	21
Figure 16 Observed and synthetic values, k-NN imputation method for ozone values from May 5th, 2012 to May 10th, 2012 in Barcelona-Palau Reial (k=5, variables considered are NO, NO ₂ , temperature, hour of the day and solar radiation).....	21
Figure 17 16 points used to calculate Jenkinson and Collison classification (meteorological team of Javier Martín-Vide, " π – Plates" project)	26
Figure 18 27 types of J&C synoptic classification over Spain (Martín-Vide et al., 2016)	26
Figure 19 Density plot of 1hO ₃	28
Figure 20 Density plot of 8hO ₃	28
Figure 21 Variation of 1hO ₃ according time variables for Barcelona – Palau Reial (PR_ZU) and Barcelona – Eixample (EI_RA).....	30
Figure 22 Variation of NO _x and solar radiation in the week	31
Figure 23 Scatter plot of every meteorological variable vs 1hO ₃ for Barcelona-Palau Reial and Barcelona-Zona Universitaria	33
Figure 24 Scatter plot of every air quality variable vs 1hO ₃ for Barcelona-Palau Reial and Barcelona-Zona Universitaria	34
Figure 25 Prequential scheme for training in Barcelona-Palau Reial and Barcelona-Zona Universitaria (entire year)	35
Figure 26 Prequential scheme for training in Barcelona-Eixample and Barcelona-El Raval (entire year)	36
Figure 27 Prequential scheme for training in Barcelona-Palau Reial and Barcelona-Zona Universitaria (days from May to September).....	36
Figure 28 Prequential scheme for training in Barcelona-Eixample and Barcelona-El Raval (days from May to September).....	37
Figure 29 Average MAE of the prequential evaluation analysis for 1hO ₃ in PR_ZU with data of the whole year using several combinations of <i>n_{tree}</i> and <i>m_{try}</i>	40

Figure 30 Out-of-bag error (RMSE.OOB) as a function of number trees for the training process of 1hO ₃ model of PR_ZU	41
Figure 31 Scattered plot Observed vs Predicted (1hO ₃ and 8hO ₃) of the whole year for Barcelona – Palau Reial and Barcelona–Zona Universitaria with the testing set	42
Figure 32 Scattered plot Observed vs Predicted (1hO ₃ and 8hO ₃) of the whole year for Barcelona–Eixample and Barcelona–El Raval with the testing set	43
Figure 33 Time series of 1hO ₃ and 8hO ₃ for the whole year in PR_ZU with the testing set ..	43
Figure 34 Time series of 1hO ₃ and 8hO ₃ for the whole year in EI_RA with the testing set ...	44
Figure 35 Variable importance for 1hO ₃ and 8hO ₃ in RF regression models of the whole year in PR_ZU.....	45
Figure 36 Variable importance for 1hO ₃ and 8hO ₃ in RF regression models of the whole year in EI_RA	45
Figure 37 Partial importance of the main variables for 1hO ₃ output of whole year in PR_ZU	46
Figure 38 Partial importance of NO, NO ₂ , NO _x and their moving averages for 1hO ₃ output of whole year in PR_ZU	47
Figure 39 Variable importance for 1hO ₃ in RF model of the whole year without O ₃ in PR_ZU	48
Figure 40 1hO ₃ variation with respect to Solar Rad and O ₃ for PR_ZU using data of the whole year	48
Figure 41 Mean error rate of the prequential evaluation analysis for categorical model of 1hO ₃ in PR_ZU with data of the whole year using several combinations of <i>n_{tree}</i> and <i>m_{try}</i> ..	49
Figure 42 Variable importance for 1hO ₃ and 8hO ₃ in RF categorical models of the whole year in PR_ZU.....	52
Figure 43 Variable importance for 1hO ₃ and 8hO ₃ in RF categorical models of the whole year in EI_RA	52
Figure 44 Variable importance for 1hO ₃ in RF categorical model of the whole year without O ₃ in PR_ZU	54
Figure 45 Average MAE of the prequential evaluation analysis for 1hO ₃ in PR_ZU with data from May to September using several combinations of <i>n_{tree}</i> and <i>m_{try}</i>	55
Figure 46 Scattered plot of Observed vs Predicted (1hO ₃ and 8hO ₃) of days from May to September for Barcelona – Palau Reial and Barcelona–Zona Universitaria with the testing set	56
Figure 47 Scattered plot of Observed vs Predicted (1hO ₃ and 8hO ₃) of days from May to September for Barcelona – Eixample and Barcelona–El Raval with the testing set.....	57
Figure 48 Variable importance for 1hO ₃ in RF regression model in PR_ZU using <i>randomForest</i> function with data from May to September adding two randomly generated categorical variables	58
Figure 49 Variable importance for 1hO ₃ in RF regression model in PR_ZU using <i>cforest</i> with data from May to September adding two randomly generated categorical variables	58
Figure 50 Variable importance for 1hO ₃ and 8hO ₃ in RF regression models of days from May to September in PR_ZU	59
Figure 51 Variable importance for 1hO ₃ and 8hO ₃ in RF regression models of days from May to September in EI_RA	59
Figure 52 Partial importance of the main variables for 1hO ₃ output with data from May to September in PR_ZU	60

Figure 53 Partial importance of SC for 1hO ₃ output with data from May to September in PR_ZU.....	61
Figure 54 Variable importance for 1hO ₃ in RF regression model of days from May to September without O ₃ in PR_ZU	62
Figure 55 1hO ₃ variation with respect to Rel Hum and O ₃ for PR_ZU using data from May to September.....	62
Figure 56 Variable importance for 1hO ₃ and 8hO ₃ in RF categorical models of days from May to September in PR_ZU	65
Figure 57 Variable importance for 1hO ₃ and 8hO ₃ in RF categorical models of days from May to September in EI_RA.....	65
Figure 58 Variable importance for 1hO ₃ and 8hO ₃ in RF categorical model of days from May to September without O ₃ in PR_ZU	67

Table of tables

Table 1 Metrics to assess to accuracy of the ML models	10
Table 2 Error metrics for different values of k and variables to fill ozone levels (O ₃) from May 5 th , 2012 to May 10 th , 2012	22
Table 3 1hO ₃ categories.....	23
Table 4 8hO ₃ Categories. These categories are defined by EPA (Environmental Protection Agency) to define the Air Quality Index (AQI) in USA. (Texas Commission on Environmental Quality, 2018).....	23
Table 5 Summary of characteristics of 1hO ₃ and 8hO ₃	27
Table 6 Number of ground-level ozone values of Barcelona – Palau Reial according to categories.....	29
Table 7 Number of ground-level ozone values of Barcelona – Eixample according to categories.....	29
Table 8 Correlation coefficients between outputs and every meteorological and air quality input for both couples of stations	35
Table 9 Combination of number of <i>n_{tree}</i> and <i>m_{try}</i> taken in RF model for training	37
Table 10 MAE of every testing set in the five folds of the prequential evaluation analysis for 1hO ₃ in Barcelona-Palau Reial and Barcelona-Zona Universitaria.....	39
Table 11 Summary of the results of the prequential evaluation for the most accurate combination of parameters for every regression model of the year	40
Table 12 Error metrics of the RF models of the whole year for every output in PR_ZU and EI_RA.....	41
Table 13 Out-of-bag MSE and RMSE for 1hO ₃ and 8hO ₃ with data of the whole year	42
Table 14 Error metrics for 1hO ₃ model with and without O ₃ as variable considering the testing set for the whole year in PR_ZU	47
Table 15 Summary of the results of the prequential evaluation for the most accurate combination of parameters for every categorical model of the year	49
Table 16 Error rate of out-of-bag samples and testing set for categorical 1hO ₃ in PR_ZU and EI_RA for the model of the whole year	50
Table 17 Error rate of out-of-bag samples and testing set for categorical 8hO ₃ in PR_ZU and EI_RA for the model of the whole year	51
Table 18 Error rate of out-of-bag samples and testing set for categorical 1hO ₃ in PR_ZU for the model of the whole year without O ₃	53
Table 19 Error rate of out-of-bag samples and testing set for categorical 8hO ₃ in PR_ZU for the model of the whole year without O ₃	53
Table 20 Summary of the results of the prequential evaluation for the most accurate combination of parameters for every regression model for data from May to September	55
Table 21 Error metrics of the testing set of the RF models with data from May to September for every output in PR_ZU and EI_RA.....	56
Table 22 Out-of-bag RMSE for 1hO ₃ and 8hO ₃ with data from May to September.....	56
Table 23 Comparison of error metrics for the testing set of 1hO ₃ models with <i>randomForest</i> and <i>cforest</i> package for PR_ZU with data from May to September	58
Table 24 Error metrics for 1hO ₃ model with and without O ₃ as variable considering the testing set for days from May to September in PR_ZU	61

Table 25 Summary of the results of the prequential evaluation for the most accurate combination of parameters for every categorical model of days from May to September	63
Table 26 Error rate of out-of-bag samples and testing set for categorical 1hO ₃ in PR_ZU and EI_RA for the model of days from May to September	63
Table 27 Error rate of out-of-bag samples and testing set for categorical 8hO ₃ in PR_ZU and EI_RA for the model of days from May to September	64
Table 28 Error rate of out-of-bag samples and testing set for categorical 1hO ₃ in PR_ZU for the model of days from May to September without O ₃	66
Table 29 Error rate of out-of-bag samples and testing set for categorical 8hO ₃ in PR_ZU for the model of days from May to September without O ₃	66

Abstract

In the last decades, the interest in predicting tropospheric ozone levels (O_3) has increased due to its detrimental effect on population health and vegetation. Although certain factors such as solar radiation are well known to have an influence on ozone levels, the effect of other variables is less clear. In this study, several regression models based on the Random Forest (RF) algorithm are generated to predict the daily maximum hourly ozone concentration level ($1hO_3$) and the daily maximum 8-hours average ozone concentration level ($8hO_3$) one day ahead in Barcelona, using air quality data, meteorological data and time variables as inputs. Two versions of the model are considered: taking information from the whole year and focusing only on summer months (May to September). In addition, classification models are created, based on thresholds inspired by current regulations for both outputs. RF regression models capture the time variation of tropospheric ozone through the year and they generate accurate estimations with acceptable deviation between the observations and predictions. In general, the categorical models of $1hO_3$ show suitable and lower error rates than $8hO_3$. However, the categories, which gather the most of the tropospheric ozone values have high accuracy and the categories with few values inside them have low accuracy. Consequently, these categorical models are not useful as a tool to alert the population about a specific ozone event. The analysis of RF models shows that the tropospheric ozone level ($1hO_3$ or $8hO_3$ according to the model) of the previous day to the prediction has the strongest association to the output. The importance of other inputs varies between the models considered; while solar radiation and day of the year are the main variables after O_3 for the whole year models, relative humidity, average dew-point deficit and weekday are also relevant in the summer models.

Prediction of tropospheric ozone concentration at urban locations using machine learning algorithms. Application to Barcelona, Spain

1 Introduction and objectives

The stratospheric ozone layer covers and protects the earth from ultraviolet radiation. However, in the tropospheric layer (Figure 1), ozone (O_3) is one of the main pollutants, which is dangerous for human health (Aljanabi *et al.*, 2020). At this layer, ozone is called ground-level ozone or tropospheric ozone. Important institutions such as The World Health Organization (WHO), The European Union (EU), and the Environmental Protection Agency (EPA) have established different admissible thresholds to protect human health and the vegetation involving two important ozone measurements: the daily maximum hourly ozone concentration level ($1hO_3$), and the daily maximum 8-hours average ozone concentration level ($8hO_3$) (Krzyzanowski & Cohen, 2008; The European Parliament and the Council of the European Union, 2008; Pernak *et al.*, 2019) in the sense that the maximum values for this measurements should not be exceeded. Hence, the prediction of these two parameters is essential to fulfil the current limits and regulations especially in urban areas where ozone concentration is mainly generated and population density is high (Malinović-Miličević *et al.*, 2021).

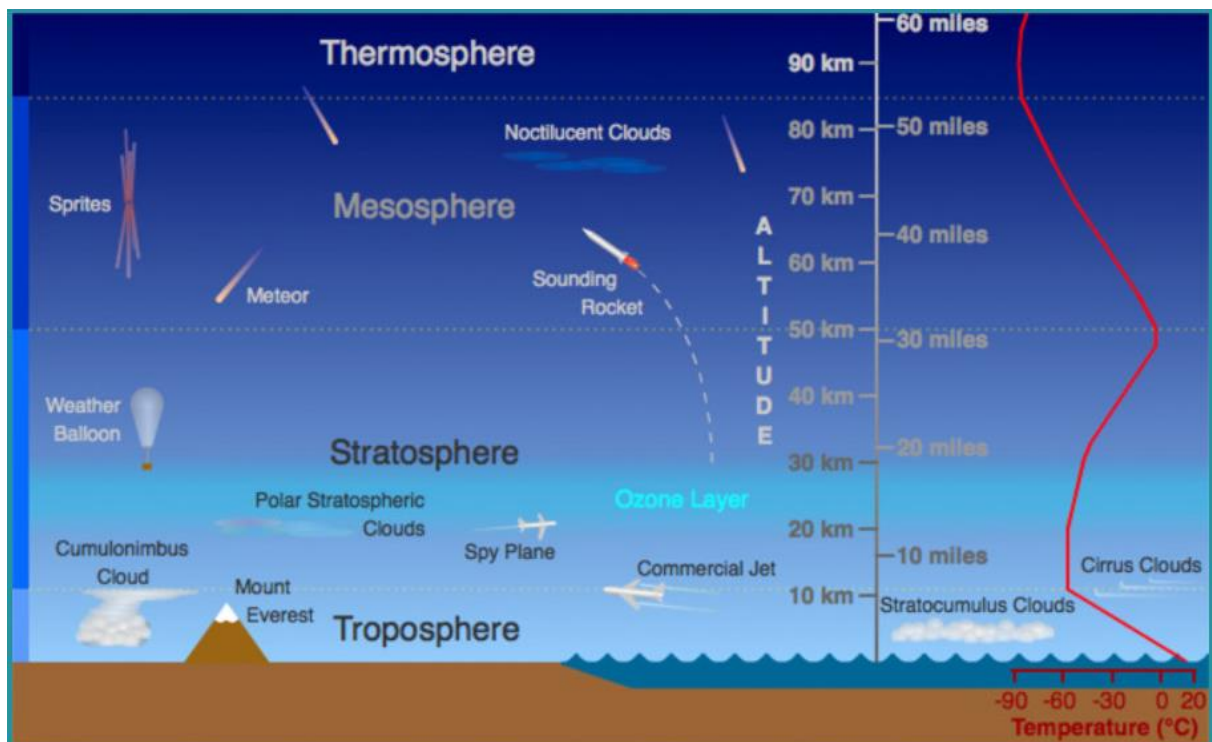


Figure 1 Atmosphere layers (UCAR Centre for Science Education, 2021)

Concerned about future tropospheric ozone levels, the government of Catalonia (Generalitat de Catalunya) along with CIMNE (International Centre for Numerical Methods in Engineering) have been working on a future model that will predict tropospheric ozone levels in Catalonia

(<https://plates.cimne.com/Reptes/OzoTroposferic>). This model is part of a large project called " $\pi - Plates$ ", which involves other models in the entire region such as the evolution of the coast lines, seismic risk areas, and flood risk areas among other. Therefore, this master's thesis is also part of the project where the prediction of tropospheric ozone is focussed on Barcelona, using air quality data and meteorological data of stations located in the city.

Machine learning (ML) is among the techniques used in the last years to predict tropospheric ozone levels (Alves *et al.*, 2019; Feng *et al.*, 2019), which has proved to produce models with suitable performances in several locations around the world. This has been possible due to the high-computational capacity available nowadays. Ozone generation at the tropospheric level involves photochemical processes and meteorological conditions (Wilson *et al.*, 2012), which complicates the prediction. That is why ML algorithms are presented as a suitable option to build a model using diverse information. Random Forest (RF) is the ML algorithm chosen for this study because of its versatility to handle several variables with a low computational cost (Breiman, 2001) and the acceptable results produced in past research (Pernak *et al.*, 2019; Meng, 2019).

This study is comprised of six chapters. In the current one, we expose the main objectives and the general steps to reach them. The second chapter involves a literature review of the main concepts and topics that we need to understand the models that we will build as well as a state-of-the-art on the use of ML algorithms to predict tropospheric ozone. In the third chapter, we introduce a description of the study area, the available data in the stations, the source of the information, and the procedure to select the time window for our study.

In the fourth chapter, every step of the methodology that we followed to build the RF models. The fifth chapter presents and explains all the results produced by the models for every period and the selected stations. The last chapter shows all the conclusions based on the results of the models and analysis of the previous chapters.

1.1 Aim

The first aim of this study is to produce machine learning (ML) models to predict the daily maximum hourly ozone concentration (1hO₃) and the daily maximum 8-hours average ozone concentration level (8hO₃) one day ahead in Barcelona, taking meteorological, air quality and time variables as inputs. It is expected that this study will help to develop an ozone prediction model for all Catalonia.

Alternatively, classification ML models will be generated also for predicting tropospheric ozone, but in the form of categorical values of 1hO₃ and 8hO₃ according to established thresholds. In other words, to predict if the values of 1hO₃ and 8hO₃ are within specific limits or categories. Finally, the photo-chemical phenomenon of tropospheric ozone generation will be analysed by interpreting the ML models and the effect of the input variables on the O₃ levels.

1.2 Overview of the methodology

The following steps were followed:

- Collect all the required data from air quality and meteorological stations.
- Pre-process the information to obtain the needed inputs and generate new features for our models.
- Divide the whole dataset into training and testing datasets, and develop the prequential evaluation analysis over the training set of values to avoid overfitting issues.
- Select the most suitable parameters for every ML model based on the error given after prequential evaluation.
- Train the ML models with the selected parameters over the whole training dataset.
- Predict the values of $1hO_3$ and $8hO_3$ using the testing data set and the trained model.
- Evaluate the results using several error metrics.
- Determine a categorical output variable of ground-level ozone concentration for $1hO_3$ and $8hO_3$ based on established thresholds.
- Train a ML model for the categorical output following the same steps employed for the previous numerical outputs ($1hO_3$ and $8hO_3$).
- Evaluate the ML categorical models based on the errors given by the confusion matrix.
- Compute and analyse the importance of input variables on the O_3 levels.

2 Literature Review

In this section, we will develop the fundamental theoretical background to understand the needed data and how machine learning algorithms work, especially focused on Random Forest.

2.1 Ground-level ozone

Also known as ‘bad ozone’, the ground-level ozone (tropospheric ozone) is the main pollutant of atmospheric smog (World Bank Group, 1998) and it has a significant harmful impact on animals and plants. Tropospheric ozone is created by photochemical reactions in the presence of sunlight between carbon monoxide (CO), nitrogen oxides (NO_x) and volatile organic compounds (VOCs) (Chameides *et al.*, 1992). There are two sources of ground-level ozone, a flux from the stratosphere and it is also generated in the troposphere (Akimoto *et al.*, 2006). In the troposphere, there are natural sources and human activities, which can cause an increase in ozone levels. Naturally, plants can emit VOCs and volcanic activity, NO_x. Gases from combustion in motor vehicles are the main source of tropospheric ozone, but also petroleum industries can produce a harmful quantity of emissions as well as energy plants or heaters at home (World Bank Group, 1998).

Ozone has a considerable impact on human health because it goes deep in the lungs causing several problems in cells of the alveoli and the most common symptoms are throat pain, coughing and irritation of the mucous (Akimoto *et al.*, 2006). It is also known that high levels of tropospheric ozone cause several damages in plants.

It is difficult to understand the phenomenon of ozone in urban areas, since several factors are involved to that. First, ozone is generated in presence of sunlight; therefore, it is also related to temperature and usually we will not have high levels of ozone during the night (Sillman, 1993). We can conclude the direct relation between Summer and high ozone levels, considering this condition. We can also understand that cities with high emissions but with not a lot of sunlight or high temperatures over the year will not have high ozone level episodes often. Second, it is something common that the highest ozone levels are registered not in the city center itself, but in places a bit far from it, which can be also related with the wind and its direction. Hence, high ozone levels can be found downwind of the city center in several urban areas (Sillman, 1993).

The Directive 2008/50/CE of European Union (The European Parliament and the Council of the European Union, 2008) establishes specific long and short term thresholds: Maximum daily 8 hour average ozone concentration of 120 µg/m³, maximum hourly ozone concentration of 180 µg/m³ for information alert and 240 µg/m³ alarm alert. These thresholds are thought to preserve human health and vegetation.

2.2 Characteristics of the ozone and meteorological stations

In order to have a wider perspective of the instruments used for these measurements, we will talk briefly about air quality and meteorological automatic stations.

2.2.1 Ozone measurements

The automatic analysers of O₃ are based on Ultra Violet (UV) absorption. O₃ is obtained due the absorption of UV rays of light at 254 nm wavelength (Haq & Schwela, 2008). Typically, the precision of these equipment is $\pm 4\mu\text{g}/\text{m}^3$.

2.2.2 Automatic meteorological stations

In this research, we have used data from automatic weather stations (AWS) in Barcelona. In these stations, measurements of a meteorological variables are made and transmitted automatically, this can be done from few variables to many of them where the information is needed (Ioannou *et al.*, 2021). In our case, we considered stations with six measurements: temperature, precipitation, solar radiation, relative humidity, wind speed, and wind direction.

There are three main parts in these stations: first, all the sensing instruments that measure the weather variables; second, the local modem, which connects the automatic station to the telecommunication system; third, a central processing system, which receives information from all the stations and it is connected to a storage system. In this way it is possible to have measurements in real time of every variable and the historical data too.

2.3 Machine learning (ML)

Since the conception of the humanity, we are surrounded by a vast amount of information. Let's take for example just the meteorological data that we discussed above, there is information about the temperature, solar radiation, wind speed, precipitation, etc. All of them are happening in this precise moment, in the past and will happen in future, and it always happened. The main difference is that now we measure these variables and can store this information; hence, we have created a huge amount of information, the most of which is available thanks to the computational capacity that nowadays exists. Machine learning (ML) is a study field, which develops computer techniques and algorithms to transform vast amount of information into suitable predicted actions and results (Lantz, 2019). Every machine learning algorithm follows five steps (Figure 2) to work properly.

We can understand that while the amount of information to analyse increases, the necessity of a higher computational capacity is bigger, and this encourages more advanced statistical methods to analyse all this information, which at the same time creates more information. This cycle allowed machine learning techniques to develop.

The information needed for the learning process (input and output data) can be shown as numerical or categorical. If it is numerical, it refers to some measurement, although if it is categorical, it refers to a specific characteristic that this information has, for example, if this data exceeds or not an established threshold.

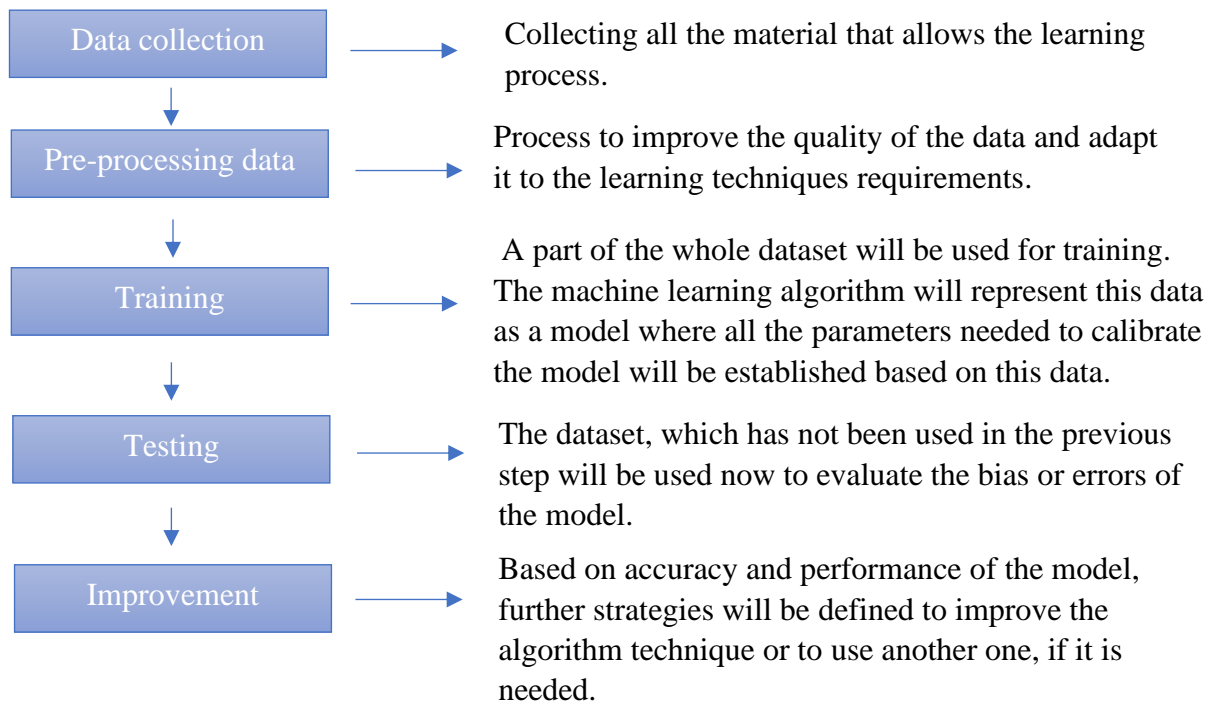


Figure 2 ML models workflow

2.4 State-of-the-art on the use of ML for predicting O₃ levels

Over recent years, machine learning (ML) algorithms have been used to predict ground-level ozone in urban areas (Pernak *et al.*, 2019; Feng *et al.*, 2019). Different inputs and outputs have been taken into account in the models as well as a different time windows and algorithms.

Abdul-Wahab & Al-Alawi (2002) developed an Artificial Neural Network (ANN) model to predict tropospheric ozone in Kuwait. Their approach was based on data from summer days and the inputs to build the model were CH₄ (methane), NMHC (non-methane hydrocarbons), CO (carbon monoxide), CO₂ (carbon dioxide), NO (nitrogen oxide), NO₂ (nitrogen dioxide), SO₂ (sulphur dioxide), temperature, wind speed, wind direction, relative humidity, solar radiation, suspended dust and tropospheric ozone from a previous measure (O₃). In this study, the output was the measure of O₃ five minutes ahead due to the lack of information of more air quality stations. In this case, the model achieved a suitable approximation to the observed values showing that relative humidity and NO as the most important variables to describe tropospheric ozone in that place.

Zabkar *et al.*, (2004) created a model using regression trees to predict the daily maximum concentration of ozone in Ljubljana-Slovenia. The particularity is that they included not only meteorological and air quality data measurement, but also data calculated by a meteorological model. Another characteristic of this study is the addition of the cosine of the day of the year as another input due to a possible approximation to the variation of ozone in the year. They did not fill missing values, but discard them and the available data was from 2002 to 2003. This researched had an acceptable approximation to the target, showing that the regression trees

model is a suitable way to predict values of maximum daily ozone level. The most important variables identified were solar radiation, temperature, and relative humidity.

Agirre *et al.*, (2007) built an ANN model to predict ground-level ozone hourly up to eight hours ahead at two stations in the Community of the Basque Country – Spain. They used hourly NO₂ and meteorological data as inputs. They also added the cosine variation of hours of the day. As we saw above, the variation of time is a variable to take into account in ozone level prediction. They used data from 2001 to 2004. The ground-level ozone at the time of prediction was the most important variable.

In Rio de Janeiro, Luna *et al.*, (2014) carried out a study to create a model using ANN to predict ground-level ozone hourly. In this case, they use eight inputs, a combination of air quality and meteorological data. Moreover, they decided to go further considering other aspects to have their set of variables, they removed rainy days, weekends, and holidays due to a variation in tropospheric ozone levels during these days. If we take into account a machine learning algorithm such as random forest (RF), which can admit many variables without increasing dramatically the computational capacity, these aspects can be considered as other variables instead of doing a pre-process of the information. Luna *et al.*, (2014) also got acceptable results predicting ground-level ozone, showing that ML algorithms are useful for this kind of task.

Pernak *et al.*, (2019) implement a model not only to predict the maximum daily 8-hours average ground-level ozone, but also to find categorical values according to some thresholds established previously, and the probability that these values of ozone exceed the thresholds. This approach was developed using RF model over data of five different locations in the Texas urban area. They included as inputs forecasting values of temperature, water vapour density, wind speed, and wind direction of the day where the ozone concentration is predicted as well as time variables such as day of the week, and day of the year; finally, maximum 8-hour average ozone concentration of the previous day. The results showed that the model can reach an acceptable prediction of ozone levels and in some specific places, a high rate of success classifying ground-level ozone according to the mentioned thresholds. Meng, (2019) also got high accurate results developing a RF model for classification of ozone days and non-ozone days in Houston, Galveston and Brazoria, USA.

Random Forest (RF) model was also used to predict hourly ozone levels and daily maximum 8-hour average ozone concentration in Hangzhou – China employing 2017 data. Feng *et al.*, (2019) calculated the variable importance of every input based on RF analysis using air quality and meteorological inputs. The results showed that NO₂ was the most important variable to predict hourly tropospheric ozone levels. However, the second most important variable was the dew-point deficit, which is related to the relative humidity and at the same more important according to this study. When they considered daily maximum 8-hour average ozone concentration (8hO₃) as output, the dew-point deficit is the most important variable. RF model showed to be suitable to predict hourly ozone and 8hO₃. As we saw, many algorithms allow computing a measure of importance of the inputs that is often useful to understand better the system under analysis and make decision. Different terms are used for this outcome, but in this text, we will use “variable importance”.

Alves *et al.*, (2019) took into account 18 inputs (air quality and meteorological data, and time variables) to build an ANN model to predict hourly tropospheric ozone levels up to 24 hours ahead in Victoria city - Brazil. They considered a sinusoidal variation of the hour of the day, the day of the week, and the month of the year. In this study unlike the others, 11 years of data were employed in the dataset. To avoid overfitting, they used the cross-validation method dividing the training dataset into 5 folds. The results of this study showed that the prediction of ozone concentration levels for the first three hours were acceptable; however, the accuracy tended to go down drastically from the sixth hour to the 24th hour.

Having a lot of inputs can lead to an inefficient system, with high computational cost. Consequently, some studies have been made developing a feature (also called inputs) selection. Aljanabi *et al.*, (2020) found that three variables were enough to obtain an accurate mean daily ozone prediction in Amman – Jordan using multi-layer perceptron neural network (MLP) and feature selection. These inputs are ozone of the previous day, temperature and humidity. In the first stage, they considered seven inputs including the day of the year and a special input, which assigned one value to a weekday and another value to a weekend day or holiday.

Malinović-Milićević *et al.*, (2021) built their model using five years of meteorological data of the previous day and the forecasted day to predict daily maximum 1-hour ozone concentration (1hO₃) and daily maximum 8-hours average ozone concentration (8hO₃) in Novi Sad – Serbia, these outputs are the same parameters taken by The Directive 2008/50/CE of European Union (The European Parliament and the Council of the European Union, 2008) to describe thresholds for tropospheric ozone. However, in this case, they considered the measurements of the forecasted day, not the result of a meteorological model as we saw previously (Zabkar *et al.*, 2004). About air quality data, they only considered as input 1hO₃ and 8hO₃ of the previous day. They also took into account time variables such as the day of the year and the day of the week for which ozone concentration level was predicted. The most important inputs according to this study were ground-level ozone of the previous day, temperature, and global radiation.

Several inputs have been employed in the models, meteorological data, air quality data, and time information. In this research, we will use all of them. Because of the large number of inputs, the capacity to add more variables without compromising computational capacity, and the acceptable results obtained previously, we have decided to build random forest (RF) models to predict 1hO₃ and 8hO₃ levels one day ahead in Barcelona.

2.5 Decision trees and recursive partitioning

Decision trees are a machine learning technique based on taking decisions making small choices at a time. They are called decision trees because they imitate the shape of a tree, where everything starts with *the root*, which can be the whole dataset. This dataset will be split, according to some criterion, ideally related to the feature with the highest predictive capability, the resulting partitions are also called *branches*. Afterwards, the partitions will be split again with another criterion, which can be related to another feature. The final result is reached in the *leaves* where the final criterion is fulfilled. The dataset is finally divided into a sufficient

homogeneous way, the tree has reached a specific size or no more partitions can be created (Nwanganga & Chapple, 2020).

Decision trees can also be used to make numeric predictions using different regression models. These models are appropriate in problems with many features and data (also called examples) to describe an output. From that point, the idea of ensembles of trees grows, where all features will be considered at once or a random group of them (Lantz, 2019). The ensemble-based method used in this research is Random Forest (Breiman, 2001).

2.6 Random forest (RF)

In order to understand Random Forest (RF), it is important to describe bagging first. Bagging is a technique, which generates new training datasets based on the original training dataset using bootstrap sampling. Bootstrapping is a method of sampling that repeats some observations and leaves behind others (Figure 3) to create a new dataset (Chong & Choo, 2011).

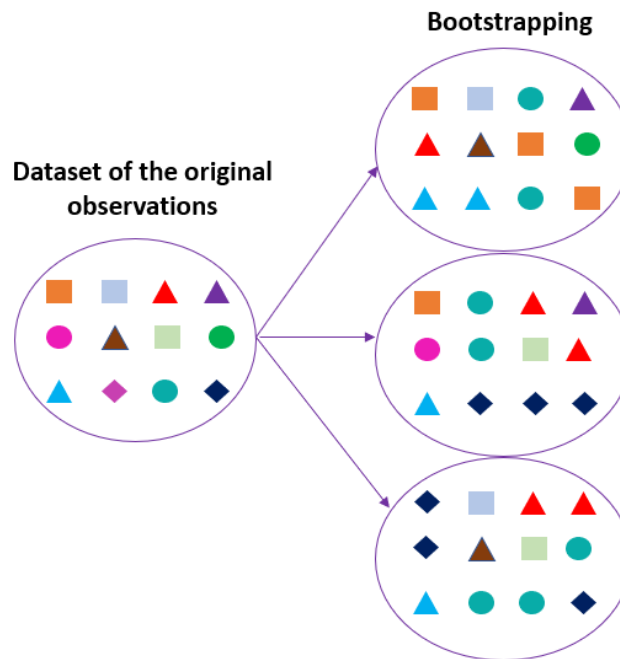


Figure 3 Scheme of Bootstrap sampling

Random forest uses bagging and combines it with a random selection of the features to grow decision trees. This means that in order to build the decision trees, only part of the inputs and bootstrap sampling from them are considered for each split. Once the ensemble of trees is created and the regression model has been applied on the decision trees, RF uses a vote to join or combine the predictions of the trees (Breiman, 2001; Lantz, 2019). The main parameters that need to be calibrated are the number of decision trees to grow (*ntree*) and the number of features to consider for every decision tree (*mtry*), the features considered for every decision tree will be selected randomly. Both parameters will be taken into account in this research, searching the most suitable combination of them for every model. Considering the principles of RF, this model allows us to employ many inputs because it takes just a part of the inputs at a time.

Breiman, (2001) mentions that in order to evaluate the performance of this model, one of the main tools is the error out-of-bag (OOB), this error is related to the data, which was not considered after bagging. In other words, this error (metric) is obtained by applying the model to the data, which does not participate in the training at that moment or it is not part of the decision tree.

2.7 Error metrics

It is essential to evaluate the accuracy of ML models (Naser & Alavi, 2020), we have seen that ML algorithms in this research will predict values of ozone. Therefore, to assess the performance of the models, we will implement several metrics, which will give us different panoramas and ideas if our model can be suitable or not. We present in Table 1 all error metrics that we will use.

Mean Error (ME)	$ME = \frac{\sum_{i=1}^n O_i - P_i}{n} \quad (1)$
Root Mean Squared Error (RMSE)	$RMSE = \sqrt{\frac{\sum_{i=1}^n (O_i - P_i)^2}{n}} \quad (2)$
Mean Absolute Error (MAE)	$MAE = \frac{\sum_{i=1}^n O_i - P_i }{n} \quad (3)$
Mean Percentage Error (MPE)	$MPE = \frac{\sum_{i=1}^n \frac{O_i - P_i}{O_i}}{\frac{n}{100}} \quad (4)$
Mean Absolute Percentage Error (MAPE)	$MAPE = \frac{100}{n} \sum_{i=1}^n \frac{ O_i - P_i }{ O_i } \quad (5)$

Table 1 Metrics to assess to accuracy of the ML models

From Table 1, we have that O_i refers to the observed output, P_i is the predicted value calculated with the ML model, and n is the total number of values in the sample. There are some characteristics of these metrics. ME is the simple error between observed and predicted output but it can be highly affected by negative results in the sum of the numerator. RMSE depends on the scale of the values of the output, the smaller this value is, the better will be the model. MAE is similar to ME but it has the advantage that it takes into account the error without being affected by negative differences. MPE and MAPE follow the same principle of ME and MAE respectively, but in this case, they consider percentual error (Naser & Alavi, 2020). MAE will represent our main parameter to select the more suitable parameters after the cross-validation or prequential evaluation analysis that we will see in the coming chapters.

2.8 Overfitting

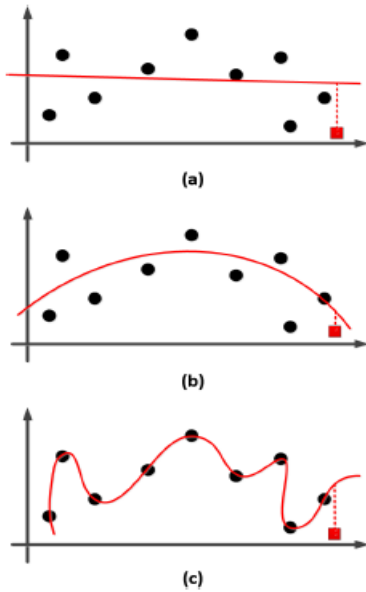


Figure 4 Underfitting, good fit and overfitting (Ghojogh and Crowley, 2019)

In section 2.3, we mentioned that the dataset needed for the ML analysis is divided into two groups, one for training and one for testing. The reason is that ML models are prone to overfitting, i.e., having a very accurate performance on the training set and a very poor or inaccurate performance on the testing set (Ying, 2019).

We can also talk about underfitting and overfitting. If the model has poor information or little, the ML model might have not enough data to learn from it. Therefore, the unseen data (or future, in red) will not be predicted suitably (Figure 4a), this is called underfitting.

In Figure 4c, we can appreciate an overfitting case, where the model is complex and has a high accuracy in the training set but a high error when this model is employed to predict the unseen value (testing dataset). Finally, in Figure 4b, we have a good fit, in the training set, we have an acceptable model with a low error, but also a low error for the unseen data or prediction (Ghojogh and Crowley, 2019).

Consequently, the questions are, where to stop the complexity of the model and how to calibrate the needed parameters of it to get an acceptable approximation in the testing set. To solve these questions, cross validation or prequential methods among others are used.

2.9 Cross validation

One of the most well-known cross validation methods is called *K-folds* cross validation. Usually, this method divides randomly the dataset into K folds or partitions, one of the partitions is used for testing and the rest of the dataset is used for training (Ghojogh and Crowley, 2019), as we can see in Figure 5.

10 partitions or folds are the most common, when we consider this method (Lantz, 2019). In Figure 5, we can see that the dataset is divided into five folds. In every case, the testing set is different. The error of the ML algorithm is the average error of the model applied on the testing set of every fold. In this way, we have a close approximation to the accuracy of the model for future predictions avoiding overfitting.



Figure 5 K-folds cross validation

2.10 Prequential evaluation method

Basic cross validation is not adequate for our study, because it does not respect the sequence of the data. Since time variables are involved, the testing set shall always be more recent than the training set.

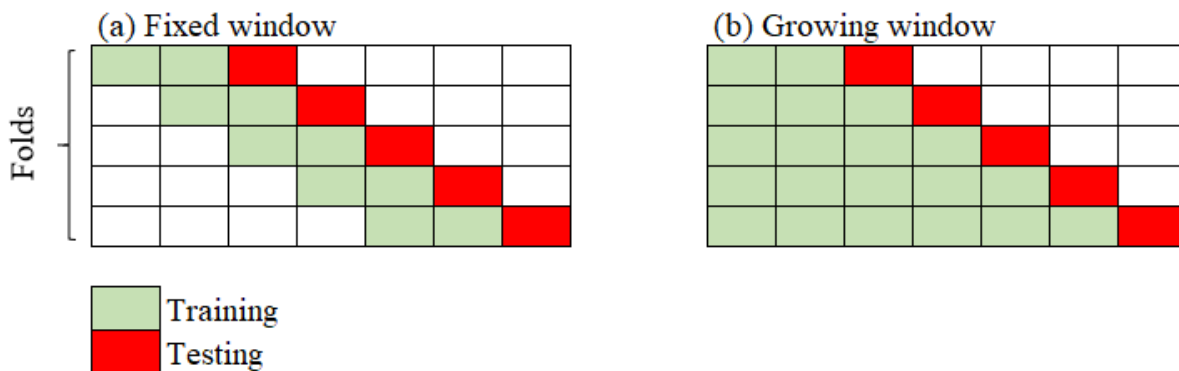


Figure 6 Schemes of Prequential evaluation method

Prequential evaluation divides the dataset into blocks as is shown in Figure 6, where the training set is always chronologically previous to the testing set (Cerqueira *et al.*, 2020). In this way, this method respects the sequential order. We can highlight two methodologies, first, fixed window, the training and testing sets move one block forward in every fold ignoring one block behind (Figure 6 (a)). This analysis was taken into account for the regression methods (numerical output).

The second methodology is called growing window and is shown in Figure 6 (b) where the training set grows always considering the past blocks, in that way, the previous information is used; however, the proportionality between training and testing sets changes (Oliveira *et al.*, 2021). We used this approach in our categorical methods (categorical outputs) in this study

because we might not have enough outputs that belong to a specific category if we ignore past information in the training analysis.

2.11 Categorical models and confusion matrix

In classification problems, the output is categorized according to some classes or descriptions. Then, the model is evaluated according to its capacity to predict if the output will be inside the right category or not. ML models can process only numerical values; therefore, every category is represented by a number (Potdar *et al.*, 2017).

Confusion matrix is the main mechanism to evaluate the accuracy of classification ML models. The rows of the matrix represent the actual class or observed and the columns represent the predicted class (Xu *et al.*, 2020).

		Predicted Class		
		A	B	C
Actual class	A	R	W	W
	B	W	R	W
	C	W	W	R

Figure 7 Scheme of the confusion matrix

In Figure 7, we can appreciate the general scheme of the confusion matrix. We have three classes, A, B and C. In the principal diagonal of the matrix, we have the number of correct predictions for each class (R), this means that both predicted and observed match. In every other cell out of the principal diagonal, we have wrong predictions (W) for every class (Lantz, 2019). The prediction accuracy can be calculated based on (7) and the respective error rate based on (8), where the error rate is one minus the accuracy.

$$accuracy = \frac{\sum R}{\sum R \sum W} \tag{6}$$

$$error\ rate = 1 - \frac{\sum R}{\sum R \sum W} \tag{7}$$

2.12 R libraries for RF

Every procedure in the methodology of this study was made using R programming language, and the main libraries to build the RF models as they were explained before have been *randomForest* (Liaw & Wiener, 2018) and *party* (Hothorn *et al.*, 2021). *randomForest* library and its function *randomForest* were used to build the models with data of the all year, *party* library and its function *cforest* were used to build the model using data from May to September. We can find below the syntaxis of the main functions:

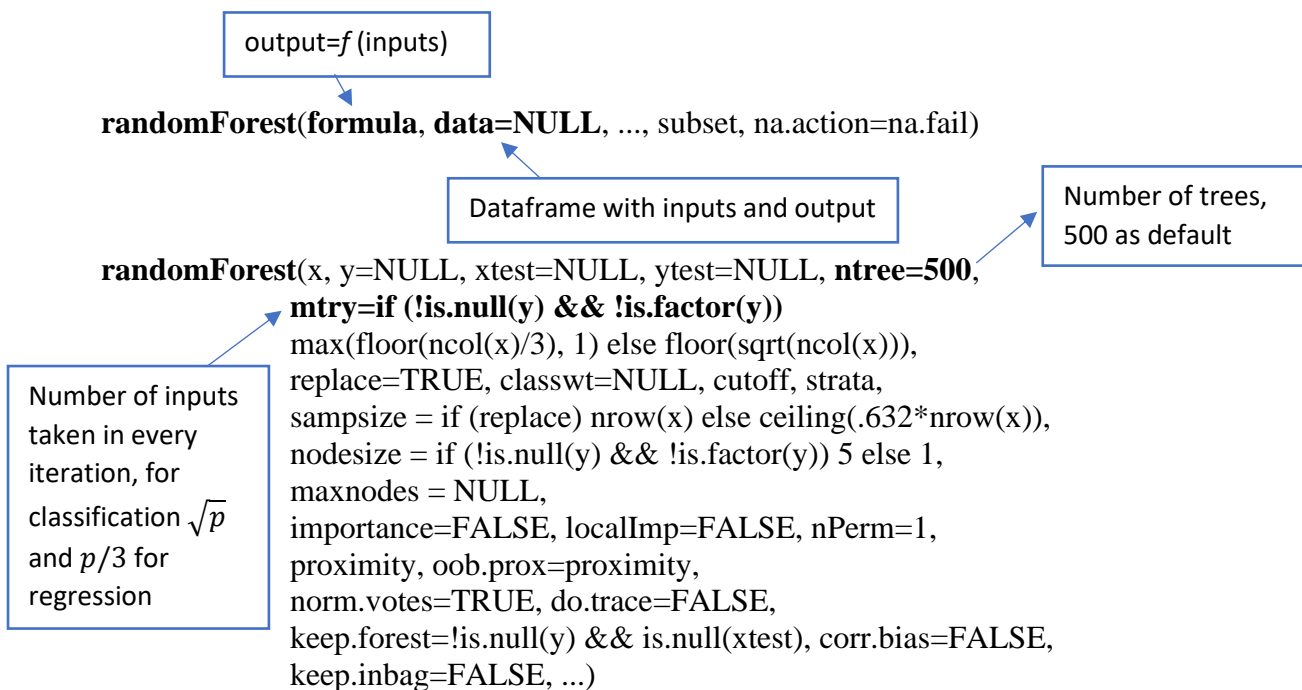


Figure 8 Main parameters of *randomForest* function

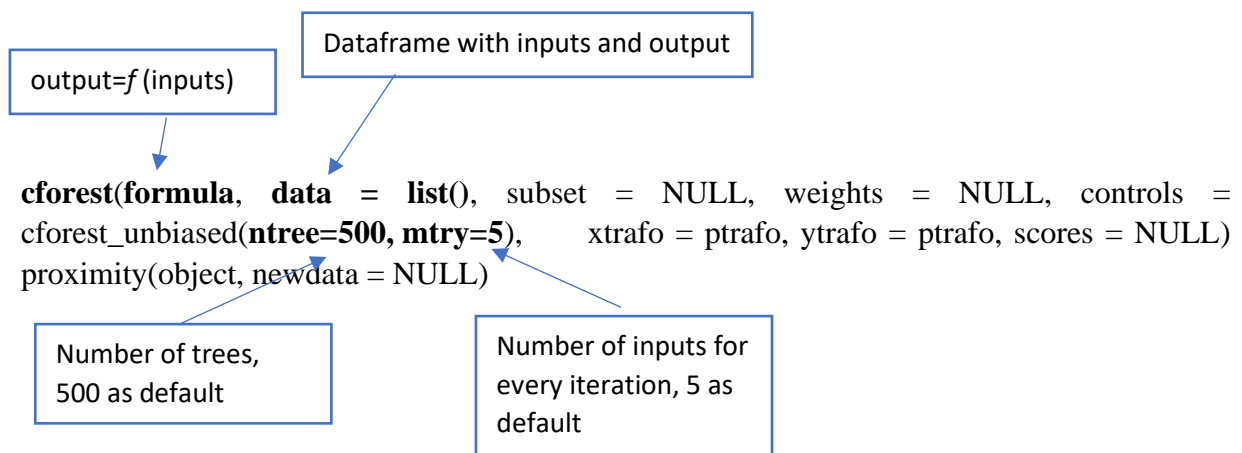


Figure 9 Main parameters of *cforest* function

In Figure 8, we see the main parameters that we will calibrate in the training and cross validation method for *randomForest* function, *ntree* and *mtry*, where p represents the number of inputs or variables. *randomForest* function follows the method that we described in section 2.6. *Formula* defines the output and the inputs, which are taken from the data frame. *cforest* and most important characteristics are shown in Figure 9, we also calibrate *ntree* and *mtry* in training process.

3 Study case

3.1 Study area

In this study, we use two couples of stations, two meteorological stations and two air quality stations, all of them located in Barcelona city – Spain. This city is located on north-eastern part of Spain, on the Mediterranean coast (Figure 10). Barcelona has a population of 1,664,182 inhabitants and a density of 16,420 inhab/Km² (IDESCAT, 2020). On the other hand, there have been some events, where pollutant substances in the air of Barcelona have exceeded the established healthy limit (SINDIC, 2019). Therefore, it is essential to forecast pollutant levels like ozone to prevent people to be exposed to health issues in this area. Another factor, which is also important is the temperature of this area, which can reach 38°C as well as solar radiation with values as high as 30 MJ/m² per day in Summer.

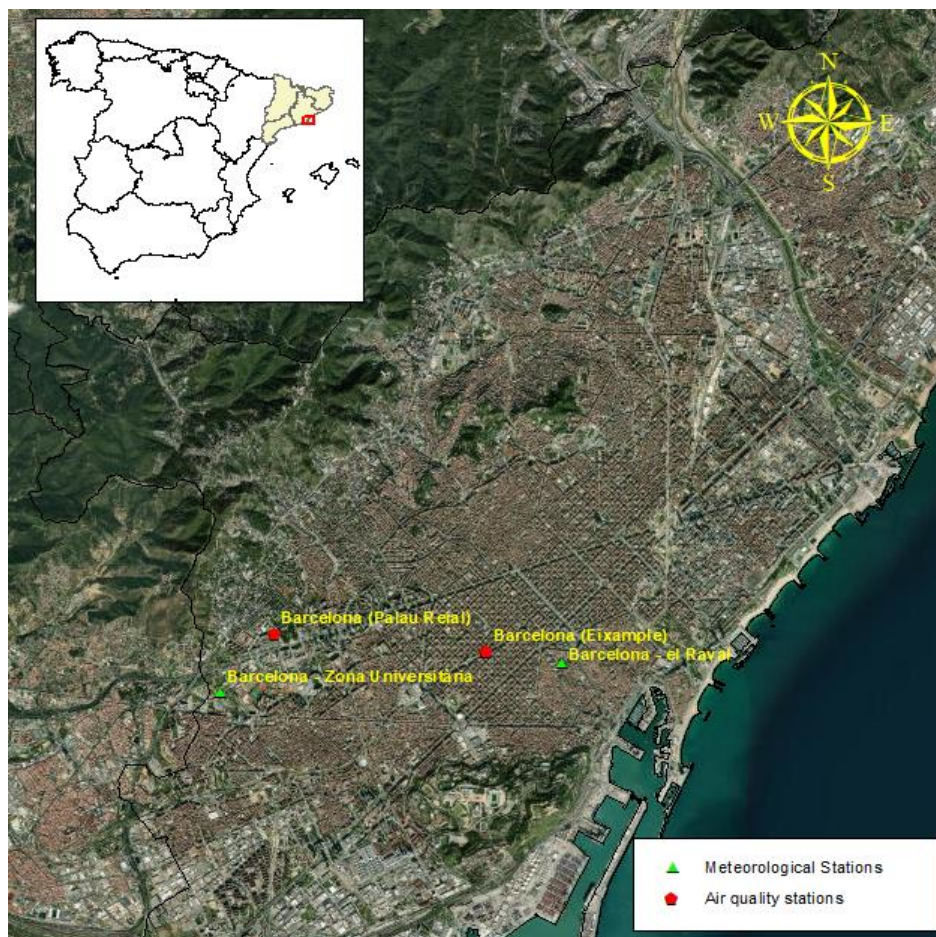


Figure 10 Location of the selected stations in Barcelona

Barcelona presents a mean temperature difference between Summer and Winter of 20°C approximately, mean annual precipitation of 600 mm, and more than 2000 sunlight hours per year with around 240 hours in Summer per month (INE, 2016). Since ground-level ozone is created in presence of sun light (section 2.1), Barcelona tends to have the majority (or all) of ozone episodes (high values of ground-level ozone) during Summer.

We mentioned in the previous section 2.1 that the main source to generate tropospheric ozone is the combustion in motor vehicles, Barcelona has an automobile fleet of 822.211 vehicles (Ajuntament de Barcelona, 2020), which means that there is a vehicle for every two people. This is an important factor to explain why the local government is concerned about the ground-ozone level, and future forecasting and alert systems. Moreover, there is a local phenomenon induced by the wind coming from the Mediterranean Sea and the surrounding mountains that creates a set of several layers of pollutants over the city (Soriano *et al.*, 2001).

3.2 Stations and data available

Dataset is comprised by two air quality stations and two meteorological stations (both of them automatic stations). In order to create a model capable to predict tropospheric ozone levels using not only air quality data, but also meteorological data, we put these stations together in two couples of stations, one of air quality data and one of meteorological data. Therefore, we have two couples of stations, first, Barcelona-Palau Reial (air quality station) and Barcelona-Zona Universitaria (meteorological station), second, Barcelona-Eixample (air quality station) and Barcelona-El Raval (meteorological station). These couples of stations are based on proximity (Figure 10), they are the closest to each other available stations. The distance between Barcelo-Palau Reial and Barcelona-Zona Universitaria (PR_ZU) is 1232 meters and between Barcelona-Eixample and Barcelona-El Raval (EI_RA) is 1177 meters. The data of each couple will be used to build a model to predict ground-level ozone. We have hourly (meteorological and air quality stations) and daily (meteorological) data available.

Based on literature review of previous research (section 2.4) we decided to take into account the following meteorological and air quality data from the stations:

Meteorological

- Temperature
- Relative humidity
- Solar radiation
- Wind speed (vector)
- Wind direction (vector)
- Precipitation

Air quality

- Carbon monoxide (CO)
- Nitrogen oxide (NO)
- Nitrogen dioxide (NO₂)
- Nitrogen oxides (NO_x)
- Ozone (O₃)
- Sulphur dioxide (SO₂)
- Particulate Matter, 10 micrometers and smaller (PM₁₀)

3.2.1 Source

Both air quality and meteorological data can be obtained from websites of the official agencies. However, we directly obtained the meteorological data from *Meteocat* (Meteorological Service of Catalonia). In case of the air quality data (hourly), we downloaded the information from *dades obertes* of Environment (Medioambient) portal of *Generalitat de Catalunya* (<https://analisi.transparenciacatalunya.cat/Medi-Ambient/Qualitat-de-l-aire-als-punts-de-mesurament-autom-t/tasf-thgu>).

3.2.2 Time window for Barcelona-Palau Reial and Barcelona-Zona Universitaria (PR_ZU)

Once we picked the stations, it is necessary to select a proper time window, where we can have data of both stations (air quality and meteorological). The time series of the available data for this couple of stations is shown in Figure 11 and Figure 12. They help us to see where we have missing values, and the period where we have data for both, the air quality station and the meteorological station. Both figures show hourly data.

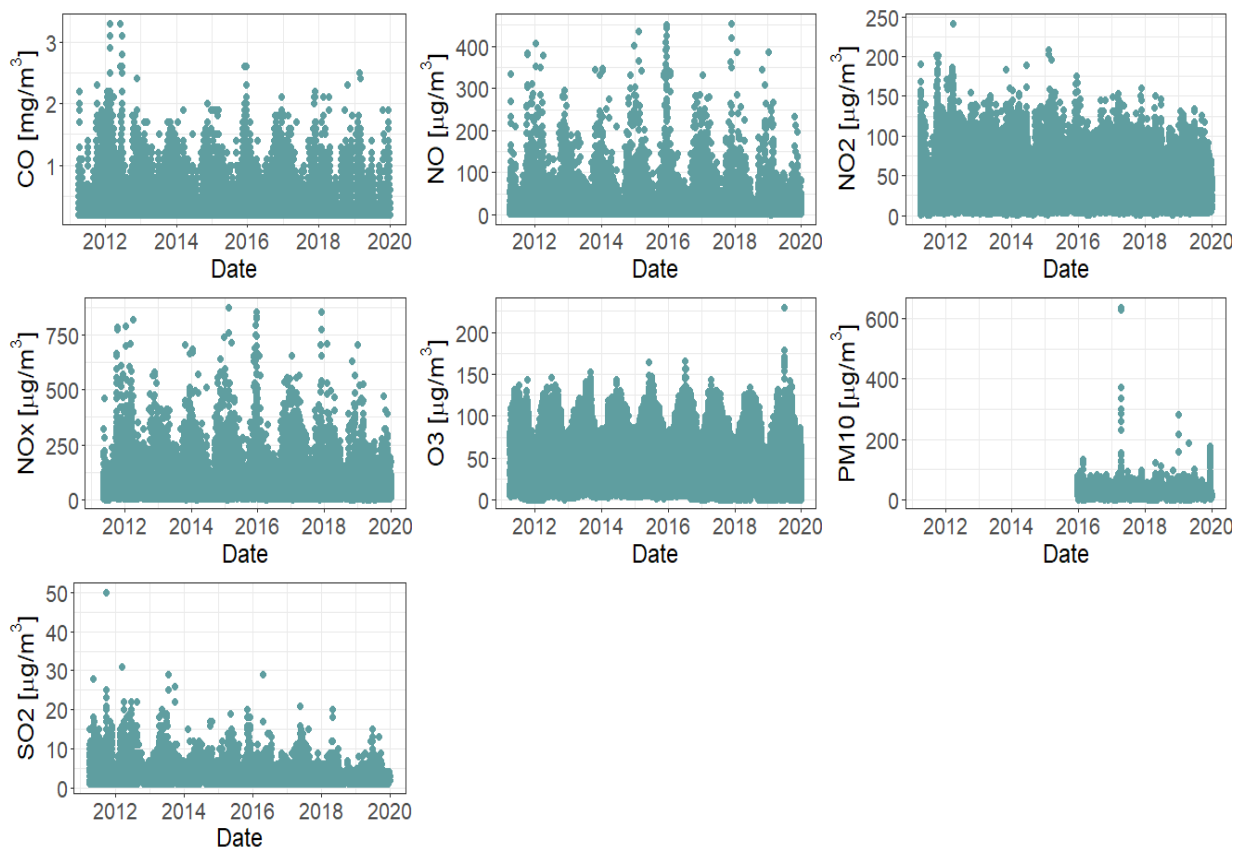


Figure 11 Time series of air quality variables data available of Barcelona-Palau Reial station

Barcelona-Palau Reial has available data from March 17th, 2011 to the present moment and Barcelona-Zona Universitaria from April 17th, 2008 to the present moment. As we can see in Figure 11, PM₁₀ has many missing years, that is why this input was not taken under consideration in the models in this case. In order to have data of both stations in the same period

of time and considering that 2020 has been atypical year (COVID-19 pandemic), we picked a time window for this couple of stations from April 1st, 2011 to December 31st, 2019.

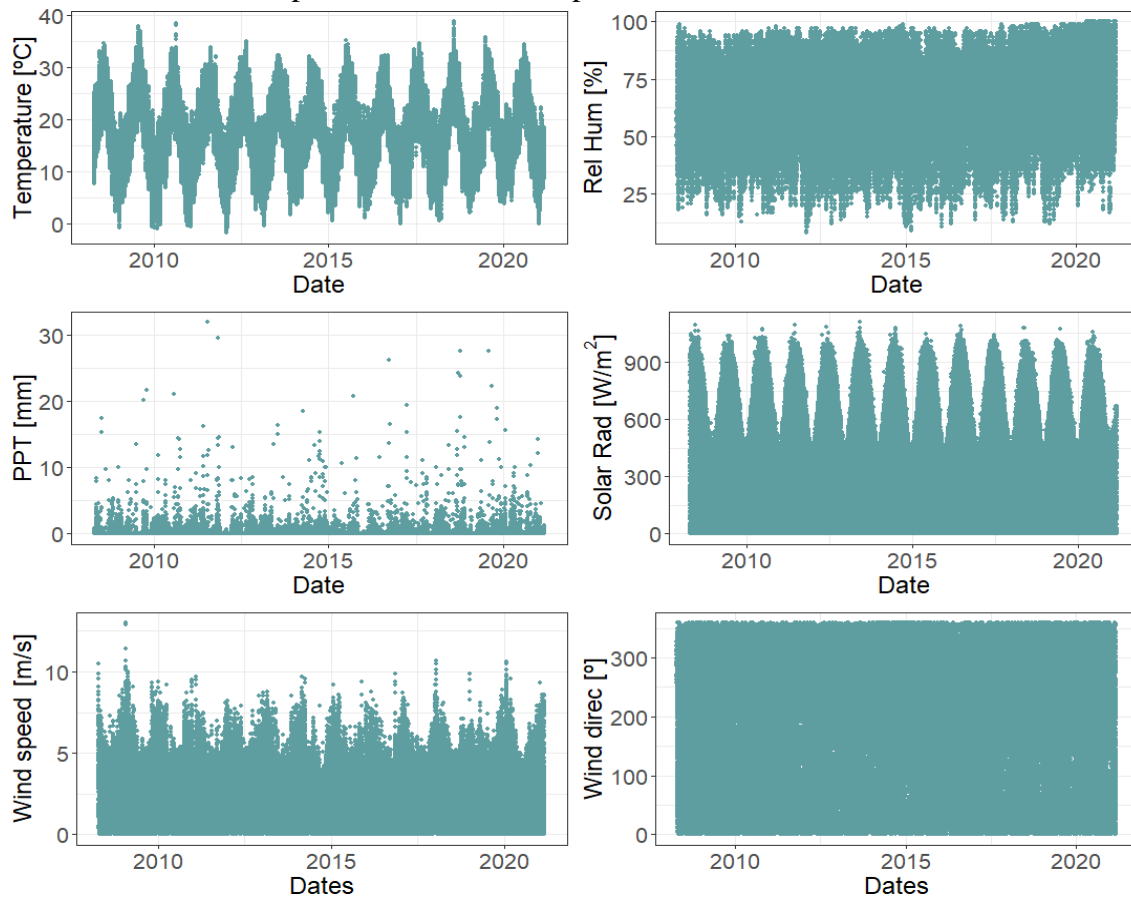


Figure 12 Time series of meteorological variables data available of Barcelona-Zona Universitaria station

3.2.3 Time window for Barcelona-Eixample and Barcelona-El Raval (EI_RA)

Hourly time series of Barcelona-Eixample and Barcelona-El Raval (EI_RA) is shown in Figure 13 and Figure 14. Barcelona-Eixample has available data of ozone levels from May 8th, 1996 to the present moment and Barcelona-El Raval from October 11th, 2006 to the present moment (Figure 14). We can appreciate in Figure 13 that there is a considerable time window with missing values from June 6th, 2009 to December 31st, 2010. Considering that situation and the necessity to have data of both stations in the same period, we selected a time window for this couple of stations from January 1st, 2011 to December 31st, 2019.

Barcelona-Eixample is catalogued as a “traffic” station, which means that it is mostly influenced by the traffic on the surrounded roads and Barcelona – Palau Reial is considered as “background” according to the source of the data (section 3.2.1), which means that it is located within an area exposed to general pollution, neither industrial nor traffic, this classification is given by the Decision 2011/850/EU (European Commission, 2013).

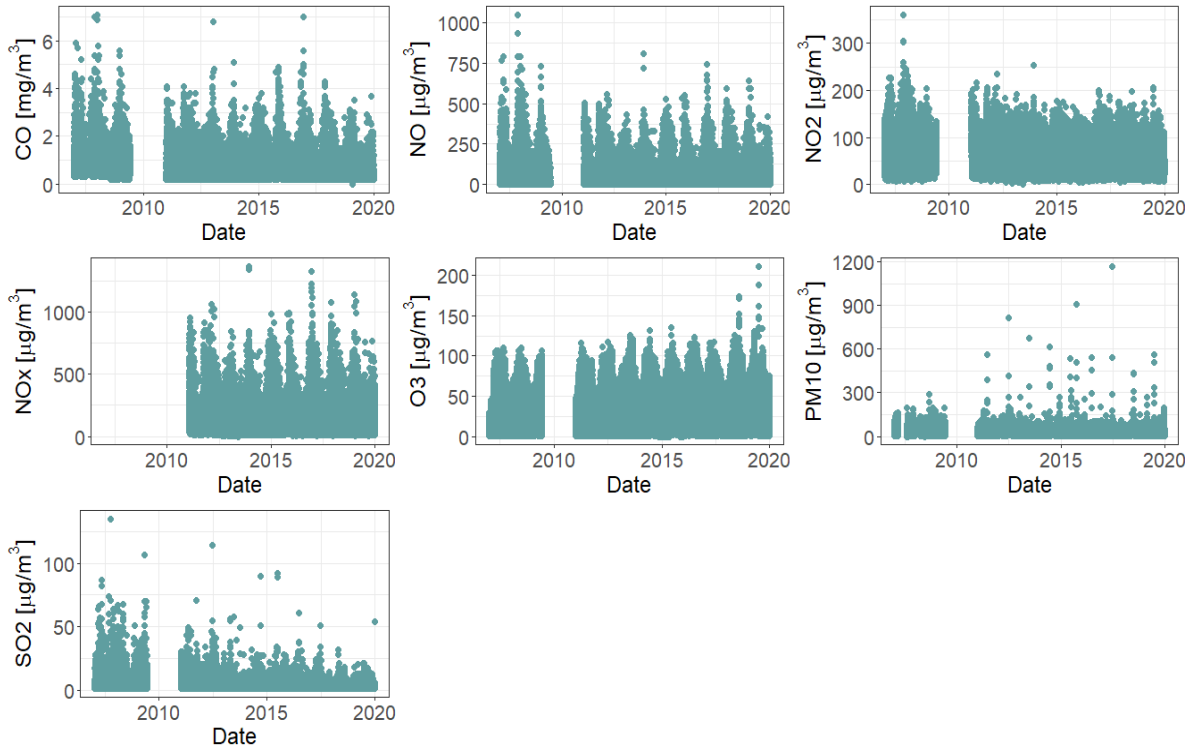


Figure 13 Time series of air quality variables data available of Barcelona-Eixample station

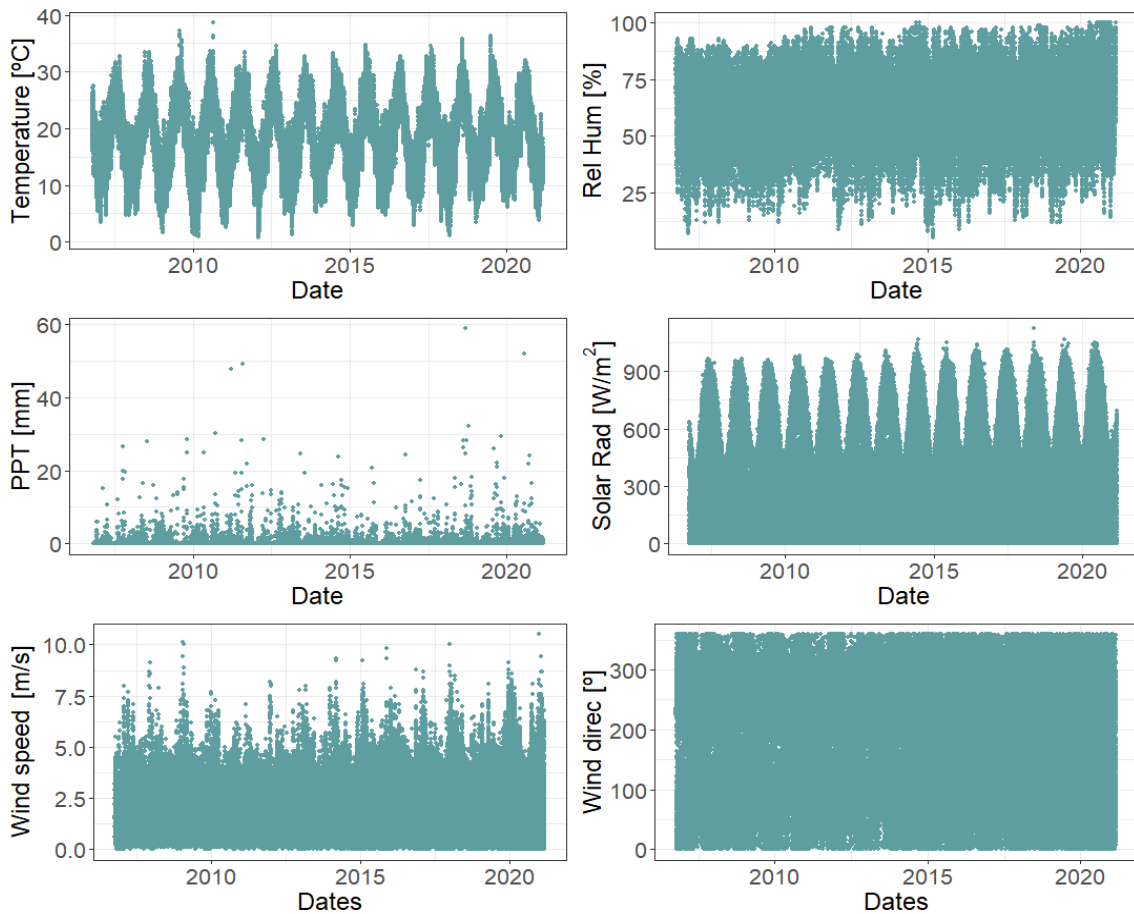


Figure 14 Time series of meteorological variables data available of Barcelona-El Raval station

4 Methodology

Figure 15 shows us the steps that we followed to develop the ML models of this research. The first step, *Data Collection* was explained in the previous section 3.2.

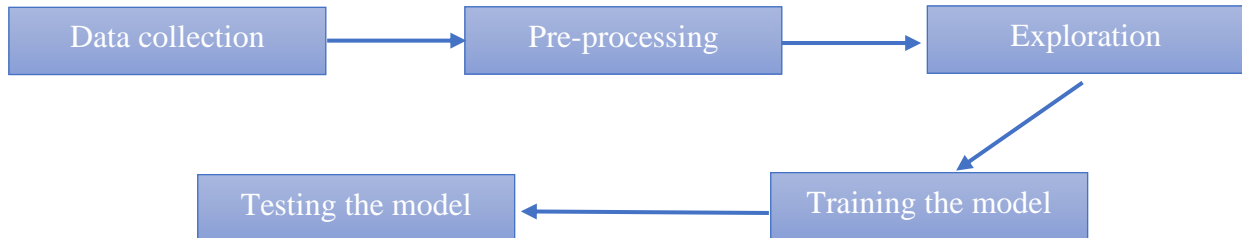


Figure 15 Steps of the methodology

4.1 Pre-processing

4.1.1 Missing values and Imputing data

There is a considerable number of missing values in the air quality data (around 14%), this situation does not occur in the meteorological data (original data); although, there are missing data, these are few (Barcelona-Zona Universitaria has four days of missing data and Barcelona-El Raval has practically no missing values). Consequently, we tried to fill missing values mainly of ozone (O_3) using k-nearest neighbours (k-NN) method (Lantz, 2019).

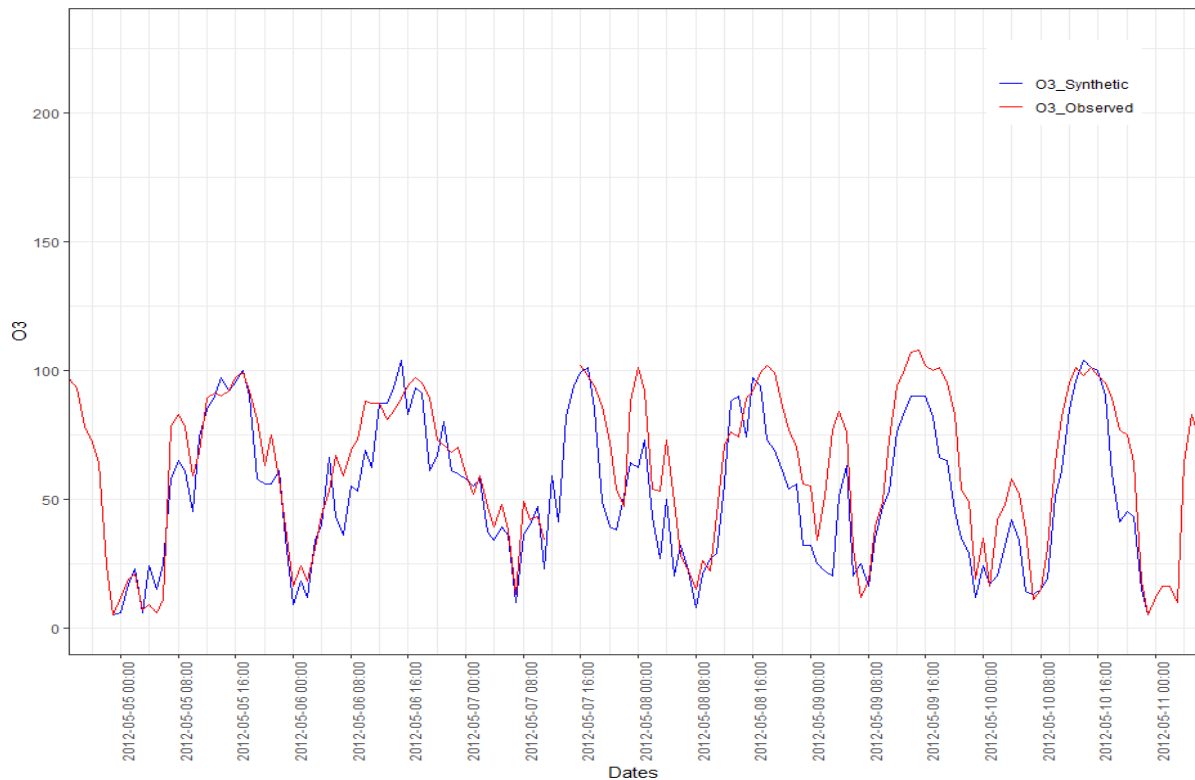


Figure 16 Observed and synthetic values, k-NN imputation method for ozone values from May 5th, 2012 to May 10th, 2012 in Barcelona-Paulu Reial (k=5, variables considered are NO, NO₂, temperature, hour of the day and solar radiation)

In Figure 16, we can appreciate an example of several attempts that we made to fill missing values using this method. In order to do build the graph shown in Figure 16, we deleted ozone data in a time window and we filled missing values using this method; finally, we compared them with the observed values. As we can see, the method can capture the tendency of ozone levels through time. However, the accuracy of the synthetic data is not high. This situation can be seen in the different attempts with different parameters and error metrics that we got in Table 2.

In this case, We have several simulations with different values of k and variables, which are part of the model to fill missing values of ozone, we appreciate that the Mean Absolute Percentage Error (MAPE, %) in the best simulation reaches 23.08% ($k = 7$; variables: NO, NO₂, NO_x, Temperature, Hours of the day and Solar radiation), which represents a considerable error and if we inserted these values to the model, we could have some inaccurate results, which will not be according to the reality. **Hence, we decided to delete missing values to run our different models.** Moreover, we consider that the number of values after deleting missing values is suitable to run the ML models (83% of the available dataset approximately).

Dates	05/05/2012		10/05/2012			
Variables: NO₂, Temperature						
k /Errors	ME	RMSE	MAE	MPE	MAPE	
2	18.09	27.6	21.89	20.75	39.55	
5	17.92	25.64	20.51	25.47	36.08	
7	17.79	25.27	20.5	24.15	36.08	
Variables: NO, NO₂, Temperature						
k /Errors	ME	RMSE	MAE	MPE	MAPE	
2	14.53	25.5	19.96	13.36	37.57	
5	15.11	24.69	20.25	15.1	36.38	
7	14.41	23.4	18.92	14.75	33.86	
Variables: NO, NO₂, NO_x, Temperature, CO						
k /Errors	ME	RMSE	MAE	MPE	MAPE	
2	13.22	25.24	20.75	7.9	39.63	
5	12.5	23.29	18.63	9.92	34.97	
7	12.67	23.17	18.66	9.46	35.43	
Variables: NO, NO₂, Temperature, Hour, Solar Radiation						
k /Errors	ME	RMSE	MAE	MPE	MAPE	
2	10.52	18.36	14.07	14.65	25.63	
5	10.31	16.79	13.05	13.21	24.71	
7	10.09	16.02	12.46	13.67	23.08	

Table 2 Error metrics for different values of k and variables to fill ozone levels (O₃) from May 5th, 2012 to May 10th, 2012

4.1.2 Outputs

We decided to take into account two numerical outputs, the daily maximum hourly ozone concentration level (1hO₃) and the daily maximum 8-hours average ozone concentration level (8hO₃) one day ahead. This decision was made based on literature review (Malinović-Miličević

et al., 2021; Feng *et al.*, 2019) and the health parameters given by WHO (Krzyzanowski & Cohen, 2008) and The Directive 2008/50/CE of European Union (The European Parliament and the Council of the European Union, 2008).

We also have categorical outputs for 1hO₃ and 8hO₃, which are explained in Table 3 and Table 4. We have two categories for 1hO₃, which are based on “unhealthy” limit given by EPA (Texas Commission on Environmental Quality, 2018); even tough, this limit is referred to 8hO₃, it allows us to have an acceptable number of values in both categories to train the model. Consequently, these categories were given to measure the capacity of the model to predict categorical values.

It was not possible to use the limits of the Directive 2008/50/CE because the corresponding thresholds of information (180µg/m³) and alert (240µg/m³) do not allow to have many values inside these limits (one in Palau Reial and two in Eixample as we can see in Figure 11 and Figure 13). Therefore, it is not possible to train the model. In Table 4, the categories were taken from the Environmental Protection Agency of USA (EPA), which are related to a healthy exposure to ground-level ozone.

Categories	1hO ₃ Range [µg/m ³]
Normal	1hO ₃ < 86
Alert	1hO ₃ ≥ 86

Table 3 1hO₃ categories

Categories	8hO ₃ Range [µg/m ³]
Good	0 ≤ 8hO ₃ < 55
Moderate	55 ≤ 8hO ₃ < 71
Unhealthy for Sensitive Groups	71 ≤ 8hO ₃ < 86
Unhealthy	8hO ₃ ≥ 86

Table 4 8hO₃ Categories. These categories are defined by EPA (Environmental Protection Agency) to define the Air Quality Index (AQI) in USA. (Texas Commission on Environmental Quality, 2018)

4.1.3 Inputs

We are taking inputs of air quality and meteorological data from the present day to predict ozone outputs one day ahead in every model:

✓ **Air quality**

1. CO (daily maximum) [mg/m³]
2. NO (daily maximum) [µg/m³]
3. NO₂ (daily maximum) [µg/m³]
4. NO_x (daily maximum) [µg/m³]
5. O₃ (daily maximum or daily maximum 8-h moving average depending on the output) [µg/m³]

6. SO₂ (daily maximum) [$\mu\text{g}/\text{m}^3$]
7. PM₁₀ (only in Barcelona-Eixample and Barcelona-El Raval) [$\mu\text{g}/\text{m}^3$]
8. Moving average of NO_x for 3 days (NO_{x_3}) [$\mu\text{g}/\text{m}^3$]
9. Moving average of NO_x for 7 days (NO_{x_7}) [$\mu\text{g}/\text{m}^3$]
10. Moving average of NO for 3 days (NO_3) [$\mu\text{g}/\text{m}^3$]
11. Moving average of NO for 7 days (NO_7) [$\mu\text{g}/\text{m}^3$]
12. Moving average of NO₂ for 3 days (NO_{2_3}) [$\mu\text{g}/\text{m}^3$]
13. Moving average of NO₂ for 7 days (NO_{2_7}) [$\mu\text{g}/\text{m}^3$]

✓ **Meteorological**

14. Daily mean temperature (Temp) [°C]
15. Maximum daily temperature (Max Temp) [°C]
16. Daily Mean Relative Humidity (Rel Hum) [%]
17. Daily Solar Radiation (Solar Rad) [MJ/m²]
18. Daily mean wind speed, vector (Wind Speed) [m/s]
19. Daily mean wind direction, vector (Wind Direct) [°]
20. Daily precipitation (Precip) [mm]
21. Maximum daily dew-point deficit (Max DPD) [°C]
22. Daily average dew-point deficit (Av DPD) [°C]
23. Moving average of temperature for 7 days (Temp7) [°C]
24. Moving average of temperature for 30 days (Temp30) [°C]
25. J&C synoptic classification (only in the model from May to September)

✓ **Time variables of the predicted day**

26. Day of the year
27. Month
28. Year
29. Weekday
30. Number of the day of the whole dataset (Num Data)

We have the same set of inputs in both prediction tasks: numerical and categorical. Two approaches were considered as for the period of analysis: first we took into account the whole year and after that, we considered a time window from May to September (adding J&C synoptic classification to the inputs) because tropospheric ozone tends to be higher in summer (Akimoto *et al.*, 2006; World Bank Group, 1998). Moreover, the government of Catalonia pays a lot of attention to ozone levels during this period (Inicio de la campaña de vigilancia de ozono troposférico, 2021).

4.1.4 Procedure to obtain output and inputs

We have a dataset comprised of hourly (air quality and meteorological stations) and daily (meteorological stations) data to obtain the needed outputs and inputs. Many of the inputs were taken directly from the provided information (section 3.2.1) such as daily mean temperature, maximum daily temperature, daily mean relative humidity, daily solar radiation, daily mean wind speed, daily mean wind direction, and daily precipitation. However, there are some inputs

that need a pre-process. Consequently, we developed different procedures according to the variable that we wanted to obtain.

4.1.4.1 Maximum daily value for air quality data

Taking into account that we have several missing values in this data, we considered a limit of 25% of missing values (Malinović-Milićević *et al.*, 2021) for every day (maximum six hours) to take the data of the day as valid, otherwise the value of the day is discarded. Every daily value fulfils this criterion. A daily value is taken into account from 0:00 to 23:59 of the specific date.

4.1.4.2 Daily maximum 8-hours average ozone concentration

To compute 8hO₃, we consider the preceding hours to the hour that we want to get this value (seven hours before in this case) and we also take into account 25% of missing values; therefore, a maximum of two hours for missing values, if this condition is not fulfilled, the value is removed. The daily value of the maximum average will be considered if no more than six hours in the day are missing, this is the same principle that we have described above.

4.1.4.3 Dew point deficit

If temperature goes down enough to produce dew or fog and we have saturation in the hair, we reach the dew point (Sensirion, 2006). Dew point can be calculated using the Magnus formula based on temperature and relative humidity:

$$Dp = \frac{\lambda \left(\ln \left(\frac{RH}{100} \right) + \frac{\beta \cdot T}{\lambda + T} \right)}{\beta - \left(\ln \left(\frac{RH}{100} \right) + \frac{\beta \cdot T}{\lambda + T} \right)} \quad (8)$$

From -45°C to 60°C:

$$\lambda = 243.12^{\circ}\text{C}$$

$$\beta = 17.62$$

$$\alpha = 6.112 \text{ hPa}$$

$RH = \text{Relative humidity (\%)}$

$T = \text{Temperature (}^{\circ}\text{C)}$

Dew point deficit is the result of the difference between air temperature and dew point. Consequently, we calculated a deficit dew point (DPD) for every hourly temperature of the original data of the meteorological stations and from there, we obtained an average dew point deficit (Av DPD) and a maximum dew point deficit (Max DPD).

4.1.4.4 Jenkinson and Collison synoptic classification for Catalonia (SC)

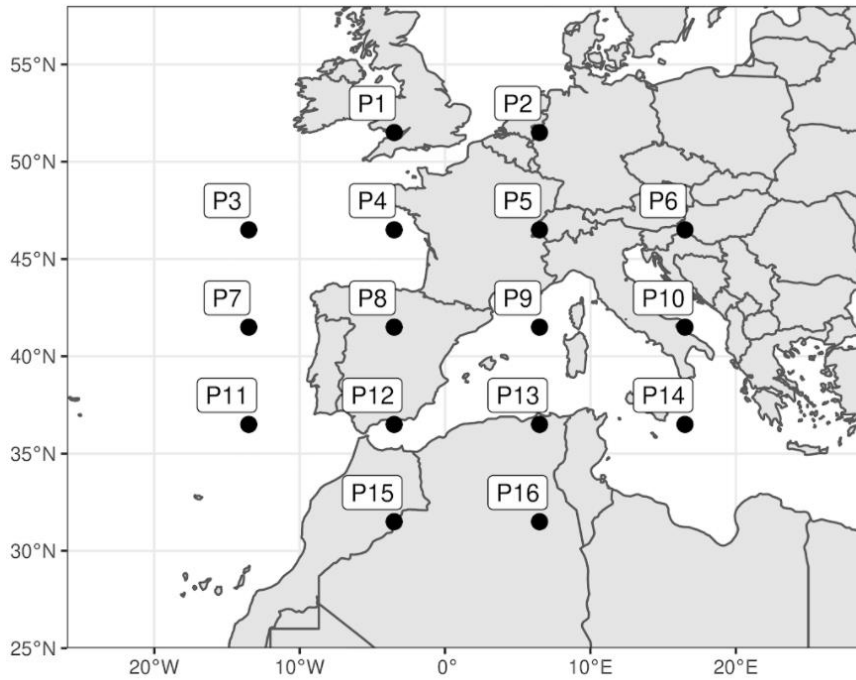


Figure 17 16 points used to calculate Jenkinson and Collison classification (meteorological team of Javier Martín-Vide, "π – Plates" project)

	<p>Pure advection in the shown directions, establishing the directions of the wind</p>
	<p>Advection with anticyclonic characteristics</p>
	<p>Advection with cyclonic characteristics</p>
	<p>A: Anticyclone C: Cyclone U: Undefined field of pressure</p>

Figure 18 27 types of J&C synoptic classification over Spain (Martín-Vide *et al.*, 2016)

A synoptic classification characterizes atmospheric circulation. Jenkinson and Collison (J&C) classification describes atmospheric circulation into 27 types obtained from 7 variables based on atmospheric pressure (Martín-Vide *et al.*, 2016). For Spain territory, and specifically in Catalonia for our study, the classification was obtained based on data of 16 different points shown in Figure 17. The schematic description of the atmospheric circulation for every type of the classification is shown in Figure 18.

The data of the J&C classification for Catalonia of every day from May to September in the period of study of both couples of stations was gotten from Javier Martín-Vide (meteorological researcher of Barcelona University) under the project "*π – Plates*" by *Generalitat de Catalunya* and CIMNE.

4.1.4.5 Moving average

The information of a single day is valid taking into account the criterion in 4.1.4.1. In air quality and meteorological inputs, we have moving averages of different data. In the case of moving average of 3 days (NO_x, NO and NO₂), we computed these values only if none of the days are missing. In the case of 7 days (NO_x, NO, NO₂ and temperature), a maximum of one day can be missing. Finally, in the case of 30 days (Temperature), a maximum of 7 missing are accepted to consider the moving average as valid.

4.2 Exploration

4.2.1 Outputs

Exploring the dataset, obtaining correlations between variables and describing them are very important steps to understand and get acceptable results when we run the ML models. In Table 5, we have the main characteristics of the outputs of the models according to the air quality stations. Ground-level ozone is higher in Barcelona – Palau Reial, this might be related to what we talked about in section 2.1 and wind has an important role to make this area more polluted in terms of ozone. We can see that the means of the values are not close to the limits given by the Directive 2008/50/CE.

Barcelona-Palau Reial					
	Minimum	Median	Mean	Maximum	Range
1hO ₃ [µg/m ³]	7.00	84.00	83.30	229.00	222.00
8hO ₃ [µg/m ³]	5.00	75.62	73.54	169.75	164.75
Barcelona-Eixample					
	Minimum	Median	Mean	Maximum	Range
1hO ₃ [µg/m ³]	3.00	64.00	63.17	211.00	208.00
8hO ₃ [µg/m ³]	1.75	54.12	52.24	143.50	141.75

Table 5 Summary of characteristics of 1hO₃ and 8hO₃

In Figure 19 and Figure 20, we can also appreciate that the ground-level ozone values of Barcelona – Palau Reial are higher than Barcelona - Eixample, and they are mainly grouped

between 25 and 125 [$\mu\text{g}/\text{m}^3$] approximately. On the other hand, 1hO₃ and 8hO₃ for Barcelona – Eixample are mostly between 12 and 100 [$\mu\text{g}/\text{m}^3$].

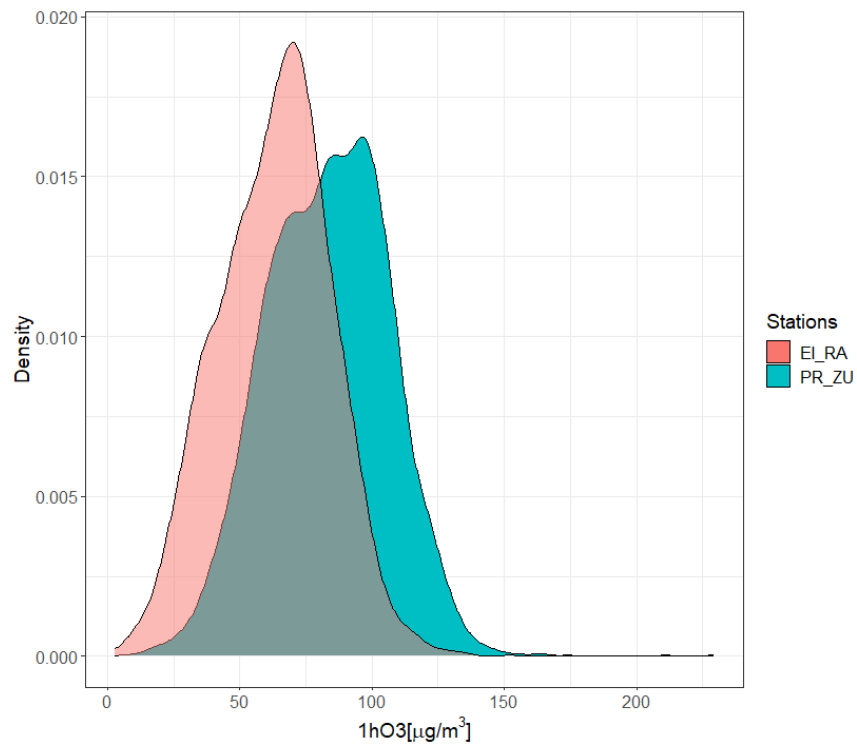


Figure 19 Density plot of 1hO₃

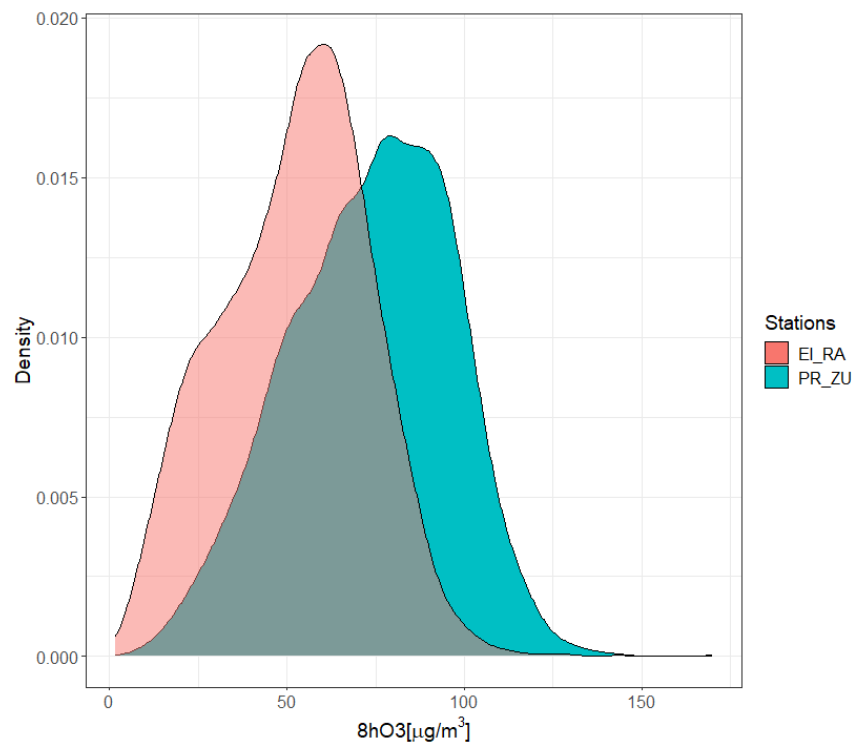


Figure 20 Density plot of 8hO₃

The values of 1hO₃ and 8hO₃ in both air quality stations vary in different ranges as we can see in Table 6 and Table 7. In Barcelona – Palau Reial, we have a similar distribution of the values

according to the categories given in section 4.1.2. However, this is not the case in Barcelona – Eixample. This can represent an issue because we will not have enough values in every category to train the categorical models in this couple of stations (Barcelona – Eixample and Barcelona-El Raval). Consequently, we might not have suitable results in this case (the number of 1hO₃ and 8hO₃ has been computed taking into account the dataset with no missing values).

We can infer that there is a considerable spatial variation of tropospheric ozone levels in the area and categorical models might not work well in every case. We did not consider quality control analysis (finding outliers and further filling of them) because it is not inside the scope of this study.

Number of 1hO ₃										
	2011	2012	2013	2014	2015	2016	2017	2018	2019	Total
< 86	98	142	170	183	145	157	142	147	149	1333
≥ 86	80	160	175	147	143	158	137	101	144	1245

Number of 8hO ₃										
< 55	39	70	60	75	81	61	54	73	48	561
≥ 55 & < 71	42	56	77	67	43	71	57	51	72	536
≥ 71 & < 86	48	64	76	98	60	56	87	55	69	613
≥ 86	48	112	132	90	104	127	80	65	104	862

Table 6 Number of ground-level ozone values of Barcelona – Palau Reial according to categories

Number of 1hO ₃										
	2011	2012	2013	2014	2015	2016	2017	2018	2019	Total
< 86	216	256	287	296	271	260	291	288	244	2409
≥ 86	29	36	40	43	34	51	49	56	46	384

Number of 8hO ₃										
< 55	166	179	160	174	162	138	176	172	146	1473
≥ 55 & < 71	50	77	109	106	95	112	99	99	86	833
≥ 71 & < 86	24	30	45	48	41	44	53	51	37	373
≥ 86	5	5	13	11	7	17	11	22	21	112

Table 7 Number of ground-level ozone values of Barcelona – Eixample according to categories

4.2.2 Trends and linear correlations between outputs and inputs

In order to show the existing trends and correlations between inputs and outputs of the model, we took 1hO₃ and all the inputs (variables) related to the prediction of it (section 4.1.3) as an example. They can be seen from Figure 21 to Figure 24.

We previously mentioned that ozone concentration levels vary seasonally, as can be seen in Figure 21 from (a) to (c), where the highest values are in summer, giving a sinusoidal variation trend to ground-level ozone through the years. The highest value of 1hO₃ took place in June. There is no considerable variation of median, minimum and maximum 1hO₃ levels over the years (Figure 21 (d)); although, there are two outstanding high levels in 2019, which are registered in both stations in the same date (29/06/2019, Saturday). Regarding weekdays, we expected to have the highest values on working days; however, higher values are registered on the weekends (Figure 21 (e)) with no significant variation on the entire week. In Figure 21 (e), the week starts on Sunday (1) and ends on Saturday (7).

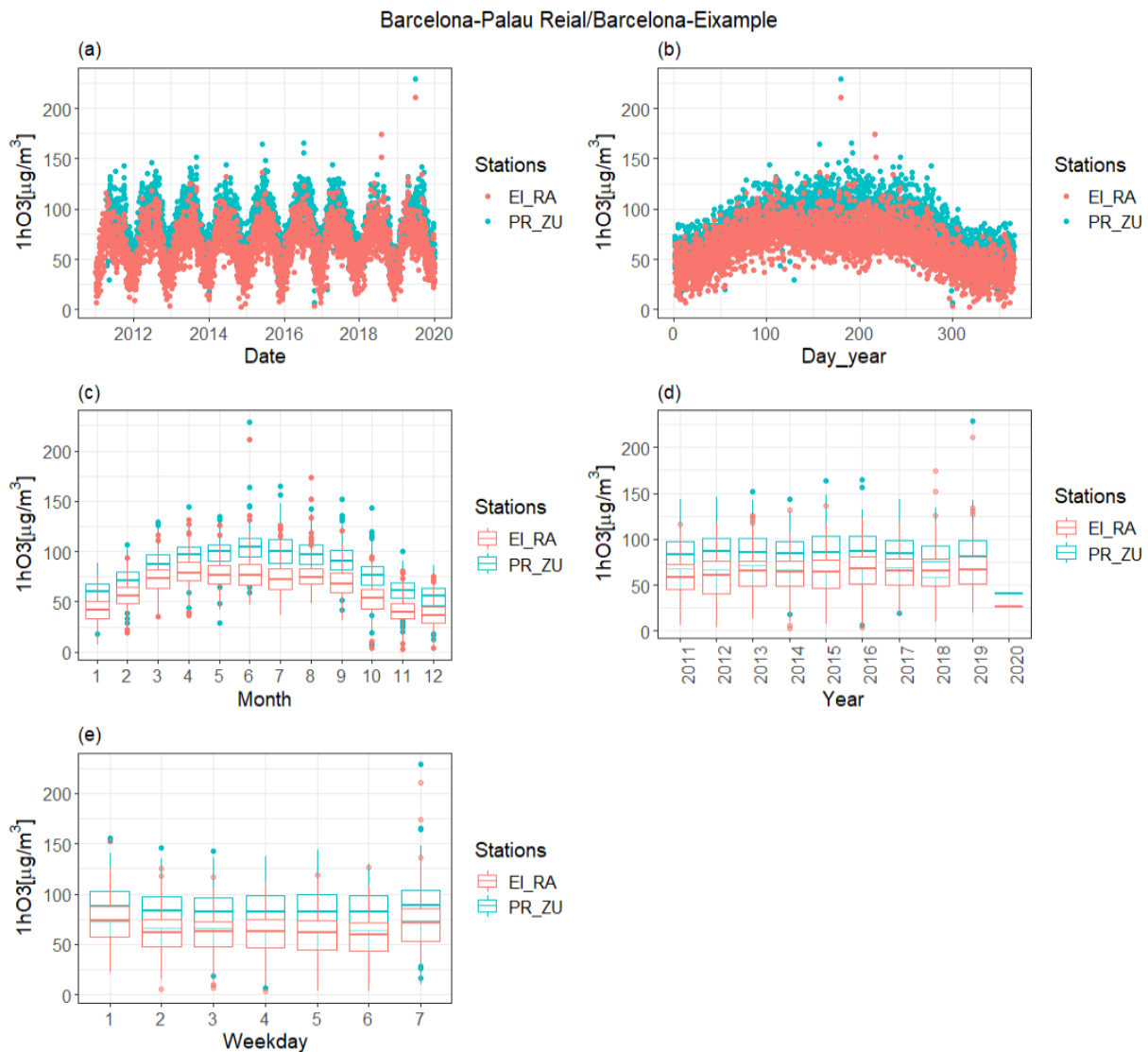


Figure 21 Variation of 1hO₃ according time variables for Barcelona – Palau Reial (PR_ZU) and Barcelona – Eixample (EI_RA)

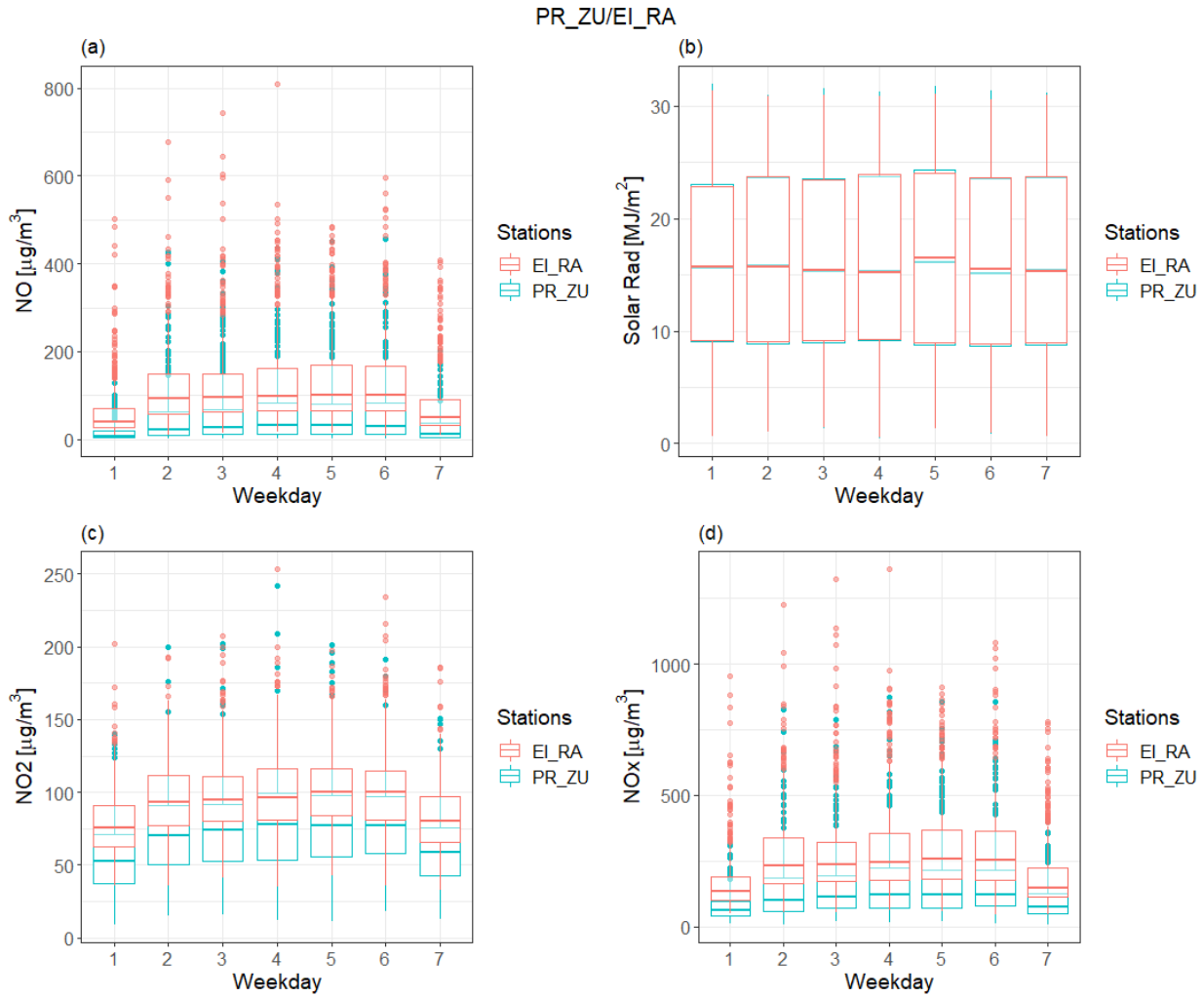


Figure 22 Variation of NOx and solar radiation in the week

We saw in section 2.1, that ground-level ozone is generated by the photochemical reaction of nitrogen oxides (NOx) and volatile organic compounds (VOCs) in presence of sunlight. Consequently, in order to search for some explanation to the variation of tropospheric ozone levels during weekdays, we also explored the variation of NO, NO₂ and NOx in weekdays along with solar radiation.

In Figure 22, we can appreciate that NO, NO₂ and NOx are higher on working days of the week than on the weekend. This is related to the gasses produced by cars and high traffic these days. Wang *et al.*, (2019) and McKeen *et al.*, (1991) showed that ground-level ozone is sensitive to changes in VOCs and NOx and the diminution of NOx might cause that O₃ levels increase. Therefore, a possible explanation for the variation of O₃ on weekdays might be related to the variation of NOx and similar components.

In Figure 23, we explore the linear correlations between all meteorological inputs and 1hO₃, we took the case of Barcelona-Palau Reial and Barcelona – Zona Universitaria as an example to see the behaviour of the variables related to the outputs. In general, we have low linear correlation coefficients (R) in every case. However, solar radiation has the highest R equal to 0.68 (Figure 23 (d)) followed by the temperature and related variables such as maximum

temperature and moving averages of this variable (Temp 7 and Temp 30), where we can see that the correlation is lower while the number of days of the moving average increases. The lowest correlation is obtained with precipitation (Precip).

As for the synoptic classification, we can see in Figure 23 (l) that U (undefined field of pressure) is the most common classification followed by A (anticyclone). However, a clear trend between high or low values of ground-level ozone and specific types of J&C synoptic classification (SC) cannot be appreciated.

Correlation between NO, NO₂, and NO_x with 1hO₃ (Figure 24) grows (inversely) while we consider more days for the moving average. From -0.27 to -0.45 for NO (Figure 24 (b) and (k)) Although, the correlations are small. In Figure 24 (e), we see that the daily maximum hourly ozone concentration level of the preceding day (O3) has a high correlation with 1hO₃ (R=0.80). In other words, the ground-level ozone of today is highly related to the ozone concentration of tomorrow. In fact, it is the variable with the highest correlation among all inputs. The correlation between the other air quality variables and the output is low in every case.

Either daily maximum hourly or daily maximum 8-hours average ozone concentration level (O3) of the previous date to the predicted output have the highest correlation coefficients for both couples of stations as we can see in Table 8, where we have a summary of correlation coefficients for every meteorological and air quality input. Solar radiation has the second highest followed by temperature; this is consistent with the definition of the creation of tropospheric ozone that we discussed previously. The same behaviour that we saw in Figure 23 and Figure 24 is repeated for all outputs and inputs, PM₁₀ is a variable only available for Barcelona-Eixample and Barcelona-El Raval (section 3.2.2); however, it shows a very low R with every output. It also important to highlight that the exploration has been made taking into account values of the whole year.

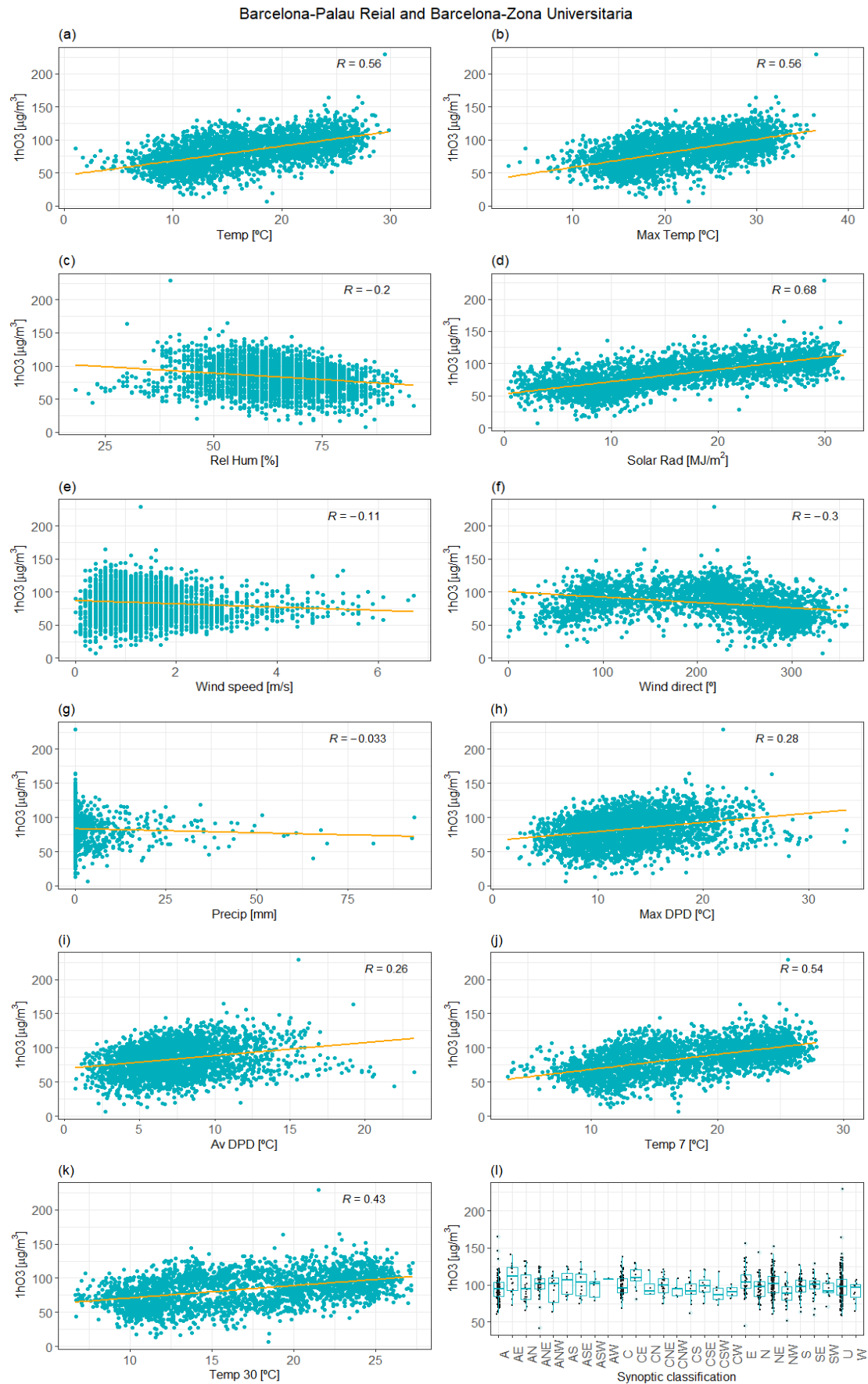


Figure 23 Scatter plot of every meteorological variable vs 1hO₃ for Barcelona-Palau Reial and Barcelona-Zona Universitaria

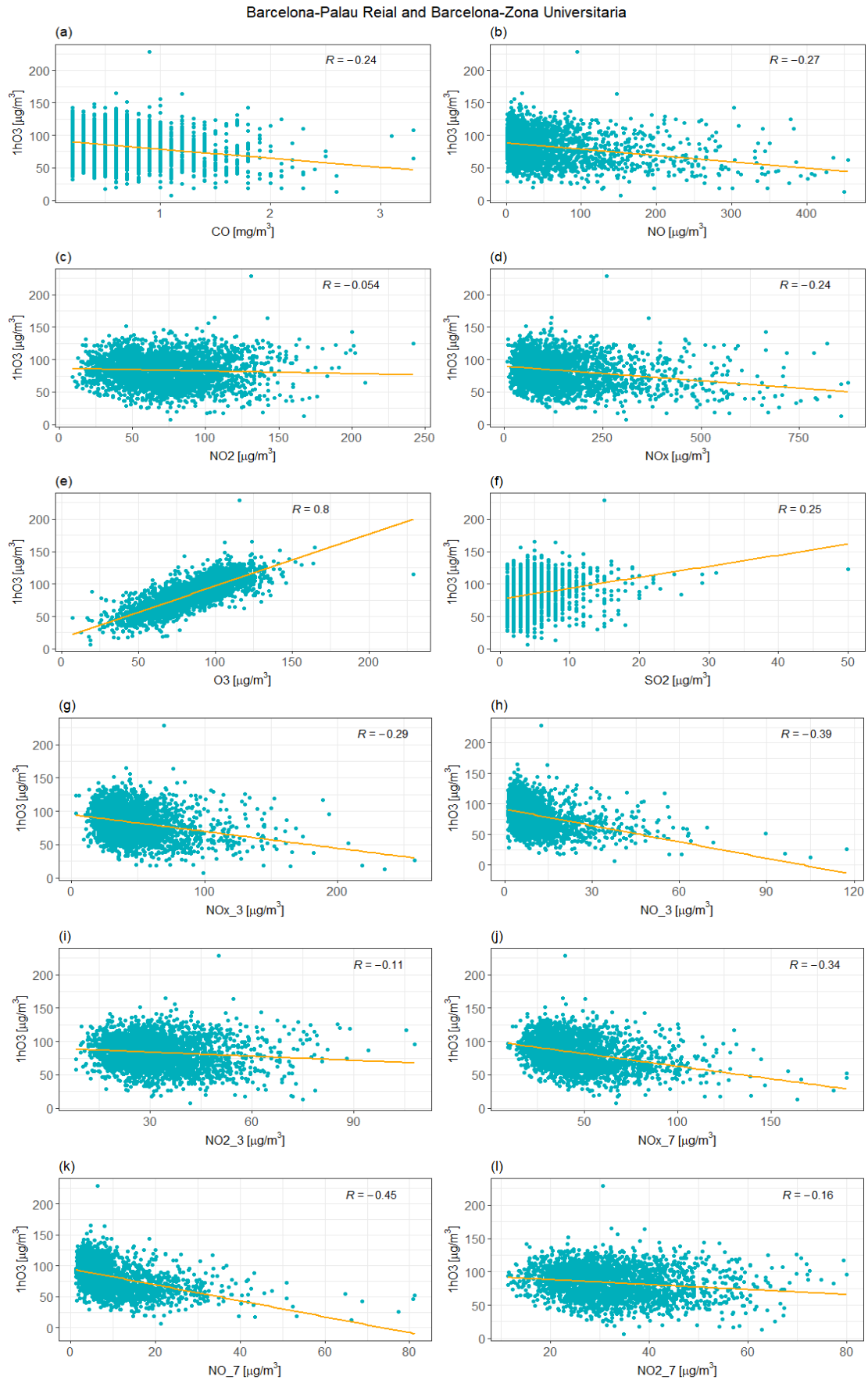


Figure 24 Scatter plot of every air quality variable vs 1hO₃ for Barcelona-Palau Reial and Barcelona-Zona Universitaria

		Barcelona-Palau Reial and Barcelona-Zona Universitaria																					
	CO	NO	NO2	NOX	O3	SO2	Temp	Max Temp	Rel Hum	Solar Rad	Wind speed	Wind direct	Precip	Max DPD	Av DPD	Temp 7	Temp 30	NOx_3	NO_3	NO2_3	NOx_7	NO_7	NO2_7
1hO ₃	-0.24	-0.27	-0.05	-0.24	0.80	0.25	0.56	0.56	-0.20	0.68	-0.11	-0.30	-0.03	0.28	0.26	0.54	0.43	-0.29	-0.39	-0.11	-0.34	-0.45	-0.16
8hO ₃	-0.30	-0.33	-0.13	-0.30	0.83	0.20	0.56	0.55	-0.21	0.69	-0.04	-0.30	-0.03	0.27	0.26	0.54	0.43	-0.34	-0.43	-0.18	-0.40	-0.50	-0.23

		Barcelona-Eixample and Barcelona-El Raval																						
	CO	NO	NO2	NOX	O3	SO2	PM10	Temp	Max Temp	Rel Hum	Solar Rad	Wind speed	Wind direct	Precip	Max DPD	Av DPD	Temp 7	Temp 30	NOx_3	NO_3	NO2_3	NOx_7	NO_7	NO2_7
1hO ₃	-0.33	-0.33	-0.09	-0.31	0.74	0.10	0.07	0.45	0.45	-0.04	0.61	0.15	-0.24	-0.03	0.14	0.09	0.42	0.31	-0.33	-0.39	-0.08	-0.42	-0.48	-0.15
8hO ₃	-0.38	-0.38	-0.15	-0.36	0.79	0.07	0.06	0.48	0.47	-0.04	0.63	0.19	-0.24	-0.03	0.13	0.09	0.45	0.34	-0.37	-0.43	-0.13	-0.46	-0.52	-0.20

Table 8 Correlation coefficients between outputs and every meteorological and air quality input for both couples of stations

4.3 Training

All available dataset has been divided into two parts, training and testing sets. Training set corresponds to the first 70% (chronologically) of all dataset and the remaining 30% for testing. In order to train the model, we applied the prequential evaluation analysis to get a suitable estimation of the accuracy of the model. We applied this analysis over the training dataset dividing this into training and testing as well.

We mentioned previously that we first, considered data of all year and second, from May to September to include J&C classification as an input. In the first case (data of all year), we took into account 5 folds for both couples of stations as is shown in Figure 25 and Figure 26.

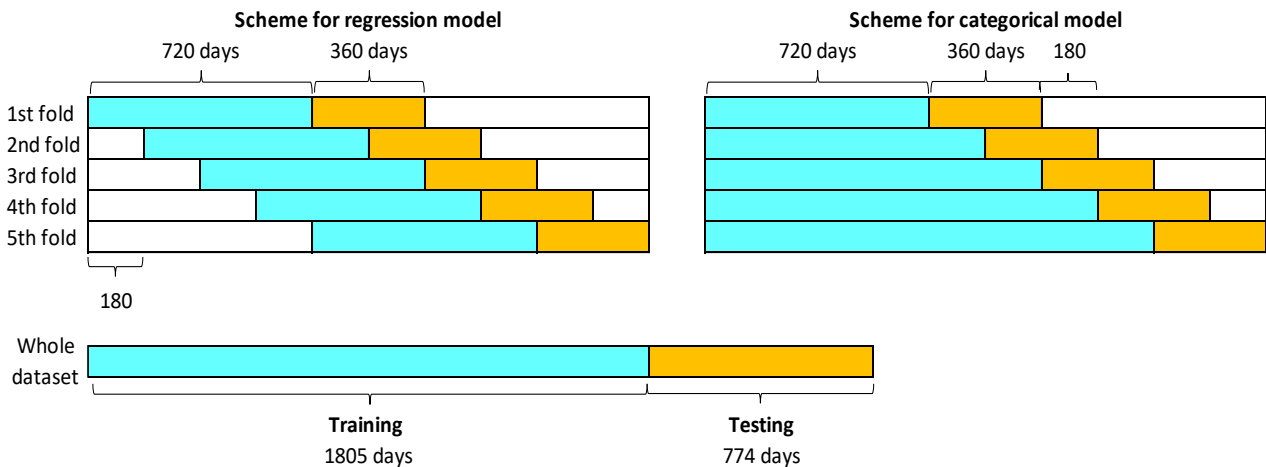


Figure 25 Prequential scheme for training in Barcelona-Palau Reial and Barcelona-Zona Universitaria (entire year)

The training set (1805 and 1956 days for PR_ZU and EI_RA respectively) is divided chronologically taking 720 days for training and 360 days for testing (the combination of both is called window). The regression model uses a fixed window and the categorical a growing one. The window moves or grows considering a block of 180 or 210 days (PR_ZU and EI_RA respectively)

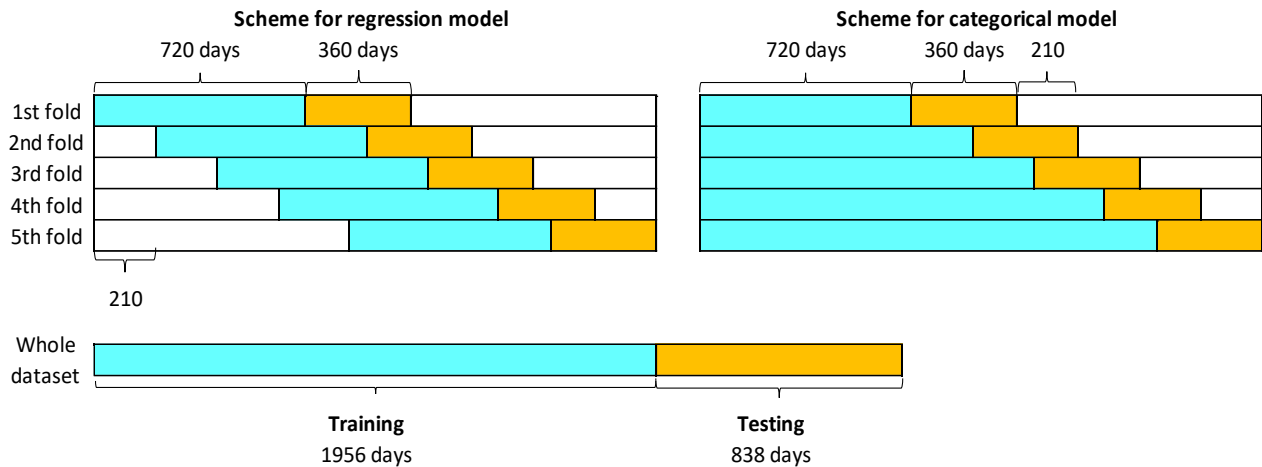


Figure 26 Prequential scheme for training in Barcelona-Eixample and Barcelona-El Raval (entire year)

For the models from May to September, we considered 3 folds based on the same prequential evaluation analysis developed previously (Figure 27 and Figure 28). Starting from 360 days for training (turquoise) and 180 days for testing (orange), the window moves or grows with a block of 110 and 120 days for PR_ZU and EI_RA respectively. We took fewer folds for this model compared to models of the entire year because we have less data available, only 773 and 783 days for training in both couples of stations respectively.

In every model, the ML algorithm is applied in the training set varying the respective parameters. In this study, we use Random Forest (RF) and for it, we calibrate the number of decision trees (*n_{tree}*) and number of features (*m_{try}*) trying several combinations, and searching for the most accurate, a summary of the combination of parameters can be seen in Table 9. In total, we considered 35 combinations of parameters. The R packages that we will use to create the RF models are *randomForest* for data of the whole year and *party* for data from May to September.

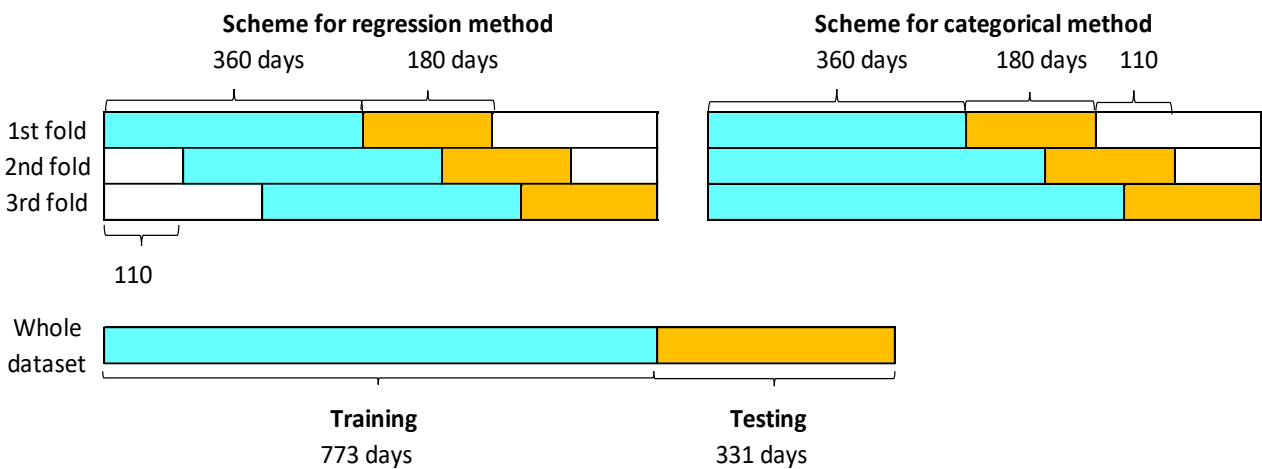


Figure 27 Prequential scheme for training in Barcelona-Palau Reial and Barcelona-Zona Universitaria (days from May to September)

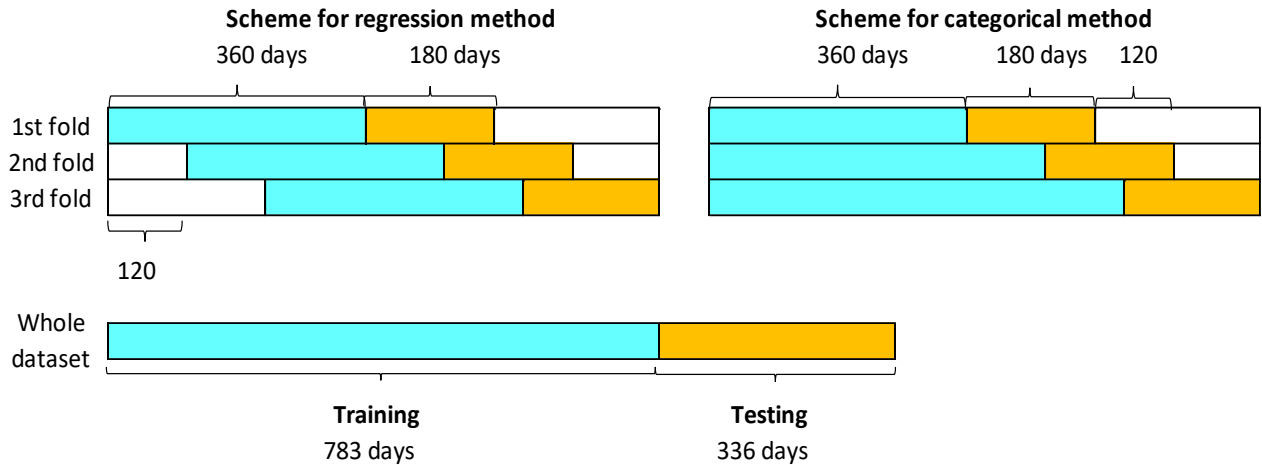


Figure 28 Prequential scheme for training in Barcelona-Eixample and Barcelona-El Raval (days from May to September)

Once the model is obtained based on the data of the training set for every combination of parameters, this model uses the inputs (features) of the testing set to predict the output, which is compared with the observed output of the testing set to calculate the mean absolute error (MAE), the average MAE computed with the result of every fold is taken as an accuracy measurement of the model with the respective parameters. The combination of parameters with the lowest MAE is taken to be applied to the whole training part of the entire dataset to obtain a new model. In the case of the categorical model, the average error rate is taken as an accuracy measurement instead of MAE.

<i>n</i> tree	200	200	200	200	200	300	300	300	300	300
<i>m</i> try	8	9	10	12	14	8	9	10	12	14
<i>n</i> tree	400	400	400	400	400	500	500	500	500	500
<i>m</i> try	8	9	10	12	14	8	9	10	12	14
<i>n</i> tree	600	600	600	600	600	700	700	700	700	700
<i>m</i> try	8	9	10	12	14	8	9	10	12	14
<i>n</i> tree	800	800	800	800	800					
<i>m</i> try	8	9	10	12	14					

Table 9 Combination of number of *n*tree and *m*try taken in RF model for training

4.4 Testing

The parameters, which gave the most accurate results (lowest MAE) in the prequential evaluation analysis are going to be used to create another model based on the data of whole training set (70% of the dataset) and the resulting model will be applied to predict ground-level ozone values based on the inputs of the testing set. Then, the predicted values will be compared with the observations of the testing set using different error metrics developed in section 2.7.

Variable importance analysis will be also developed for every RF model. The Variable importance will be computed based on mean decrease impurity (MDI) and mean decrease accuracy (MDA). Impurity is measured by residual sum of squares for regression models and by Gini index for classification models (models with categories as outputs). MDA is computed

for every variable based on the difference between the mean squared error (MSE) of OOB values and the MSE after a specific variable (the one the we want to know the MDA) is permuted. We compute the average of this difference for every tree (Liaw & Wiener, 2018).

To compute these measurements of importance, both *randomForest* and *cforest* packages have inner functions to do it. The partial importance of some specific chosen inputs will be computed as well to see how tropospheric ozone concentration varies due to these variables. The specific inputs will be considered according to their measurement of the general importance computed previously.

5 Results

Results are presented in two parts: First, all models related to the data of the whole year are exposed, second, the results of days from May to September will be taken into account.

5.1 Models of the whole year

5.1.1 Regression models

5.1.1.1 Prediction accuracy

We applied the training process described in 4.3 to calibrate the model parameters. The outputs of the models are $1hO_3$ and $8hO_3$. In Table 10 and Figure 29, we appreciate the results of the prequential evaluation analysis for $1hO_3$ after we applied the RF model to every fold for every combination of parameters.

<i>n</i> tree	<i>m</i> try	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
200	8	9.513	9.230	10.529	9.784	10.304	9.872
300	8	9.508	9.257	10.561	9.807	10.333	9.893
400	8	9.599	9.229	10.518	9.871	10.307	9.905
500	8	9.528	9.206	10.496	9.868	10.293	9.878
600	8	9.552	9.195	10.475	9.871	10.283	9.875
700	8	9.563	9.172	10.483	9.845	10.284	9.869
800	8	9.533	9.171	10.462	9.838	10.286	9.858
200	9	9.689	9.117	10.441	9.750	10.158	9.831
300	9	9.492	9.144	10.450	9.758	10.153	9.799
400	9	9.521	9.124	10.372	9.771	10.184	9.794
500	9	9.549	9.122	10.371	9.749	10.185	9.795
600	9	9.571	9.115	10.382	9.753	10.195	9.803
700	9	9.577	9.108	10.365	9.760	10.183	9.798
800	9	9.595	9.122	10.393	9.746	10.175	9.806
200	10	9.679	9.151	10.370	9.741	10.303	9.849
300	10	9.615	9.128	10.427	9.736	10.334	9.848
400	10	9.615	9.147	10.380	9.766	10.299	9.841
500	10	9.610	9.116	10.387	9.784	10.242	9.828
600	10	9.653	9.117	10.347	9.781	10.252	9.830
700	10	9.660	9.127	10.337	9.785	10.270	9.836
800	10	9.596	9.095	10.337	9.773	10.279	9.816
200	12	9.731	9.093	10.315	9.741	10.321	9.840
300	12	9.692	9.111	10.288	9.686	10.264	9.808
400	12	9.646	9.092	10.285	9.698	10.225	9.789
500	12	9.645	9.108	10.250	9.687	10.203	9.779
600	12	9.600	9.104	10.236	9.687	10.182	9.762
700	12	9.651	9.117	10.277	9.678	10.194	9.783
800	12	9.637	9.112	10.264	9.695	10.187	9.779
200	14	9.848	9.099	10.296	9.794	10.342	9.876
300	14	9.765	9.074	10.222	9.832	10.332	9.845
400	14	9.702	9.061	10.177	9.829	10.339	9.821
500	14	9.766	9.068	10.210	9.819	10.304	9.833
600	14	9.721	9.050	10.213	9.797	10.312	9.819
700	14	9.708	9.035	10.213	9.795	10.326	9.815
800	14	9.694	9.029	10.200	9.804	10.319	9.809

Table 10 MAE of every testing set in the five folds of the prequential evaluation analysis for $1hO_3$ in Barcelona-Palau Reial and Barcelona-Zona Universitaria

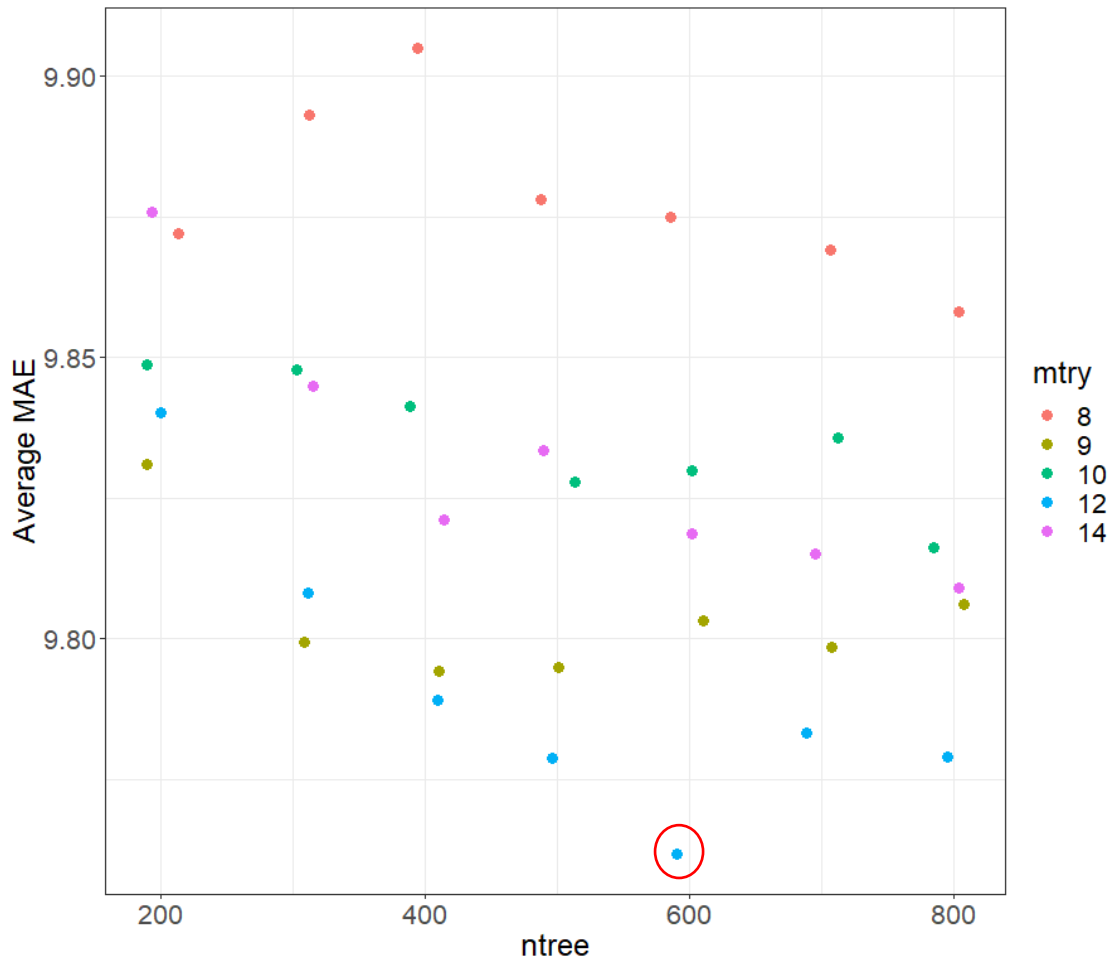


Figure 29 Average MAE of the prequential evaluation analysis for 1hO₃ in PR_ZU with data of the whole year using several combinations of *ntree* and *mtry*

The combination of parameters with the lowest mean absolute error (MAE) for RF model with 1hO₃ as output is *ntree* = 600 and *mtry* = 12 (MAE = 9.762). The same combination of parameters is applied to the whole training set (70% of the dataset) to obtain the RF model, which will be used to predict 1hO₃ values taking into account the inputs of the testing set.

A summary of the most accurate combinations of parameters for every RF regression model of the whole year with their respective outputs is given in Table 11. We can notice that MAE results are close for every model, even considering that we have different output. We applied these parameters to the whole training set in the respective models.

Output	Stations	<i>ntree</i>	<i>mtry</i>	MAE
1hO ₃	PR_ZU	600	12	9.762
8hO ₃	PR_ZU	400	9	9.571
1hO ₃	EI_RA	500	12	9.718
8hO ₃	EI_RA	700	12	9.382

Table 11 Summary of the results of the prequential evaluation for the most accurate combination of parameters for every regression model of the year

After we train the model using the parameters with the most accurate approximations to the outputs and the whole training set, we get the error metrics related to the training set and the testing set. In the first one, all error metrics are low when we compare the predicted output and the observed one because this part of the dataset was employed to build the model itself (or train). However, when we apply the inputs that belong to the testing set, the error metrics increase as it is shown in Table 12 where we have a summary of the error metrics for every model considering both training and testing sets.

			ME	RMSE	MAE	MPE	MAPE
PR_ZU	1hO ₃	Training set	0.003	4.787	3.615	-1.278	5.100
		Testing set	-3.248	13.423	9.970	-7.195	13.725
	8hO ₃	Training set	0.006	4.622	3.554	-1.932	6.272
		Testing set	-2.925	12.122	9.441	-8.486	15.816
EI_RA	1hO ₃	Training set	0.033	4.782	3.683	-2.938	7.983
		Testing set	-1.434	12.467	9.619	-7.553	18.076
	8hO ₃	Training set	0.002	4.396	3.447	-4.423	10.095
		Testing set	-1.420	11.194	8.776	-11.777	23.012

Table 12 Error metrics of the RF models of the whole year for every output in PR_ZU and EI_RA

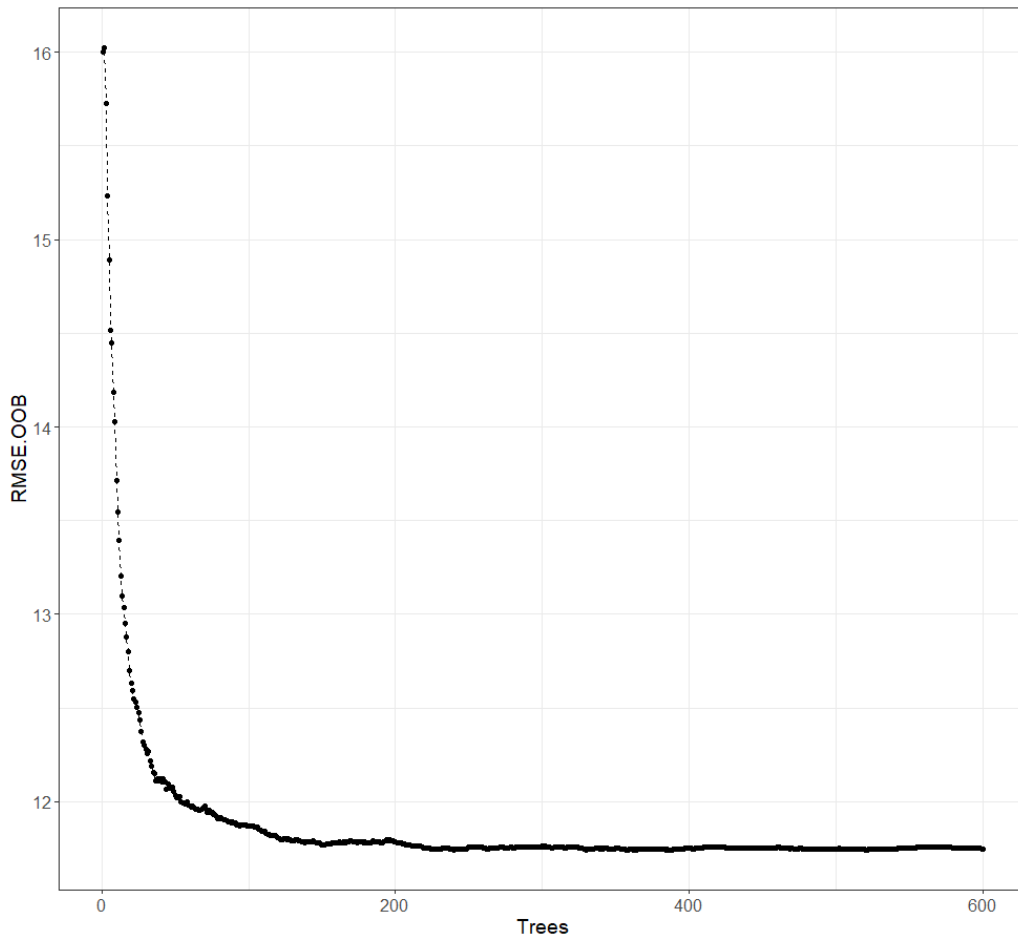


Figure 30 Out-of-bag error (RMSE.OOB) as a function of number trees for the training process of 1hO₃ model of PR_ZU

		MSE.OOB	RMSE.OOB
PR_ZU	1hO ₃	138.040	11.749
	8hO ₃	128.3307	11.328
EI_RA	1hO ₃	137.7015	11.735
	8hO ₃	116.6584	10.801

Table 13 Out-of-bag MSE and RMSE for 1hO₃ and 8hO₃ with data of the whole year

RMSE and MAE of the predicted and observed values are close in every testing set for each model. These results can be also compared with the RMSE of the out-of-bag (OOB) values obtained in the training process. In Figure 30, we can appreciate how the mean squared error (MSE) for OOB values varies according to the number of trees. The error given in the OOB analysis can be also taken into account as a suitable approximation of the accuracy of the model. Table 13 shows the MSE and RMSE of the OOB analysis (these values are related to the number of trees taken into account in every model). RMSE obtained in the analysis of the testing set and OOB values are quite similar. There is a considerable variation in the mean absolute percentage error (MAPE), from 13.7% to 23% (Table 12). MAPE indicates the average percentage difference between observed and predicted value. If the observation is equal to 1 and prediction is equal to 2, MAPE will be 100%.

There is some dispersion in the scattered plot of the observed vs. predicted of 1hO₃ and 8hO₃; however, the tendency is close to the ideal case (blue line) for both stations and both output (Figure 31 and Figure 32), there are few points separated from the general trend, which are related with events with an extraordinary high tropospheric ozone concentration. 8hO₃ model for EI_RA shows the highest dispersion among the graphs, this can also be seen in MAPE value that we mentioned before. The scattered plots were made using only the outputs after testing process.

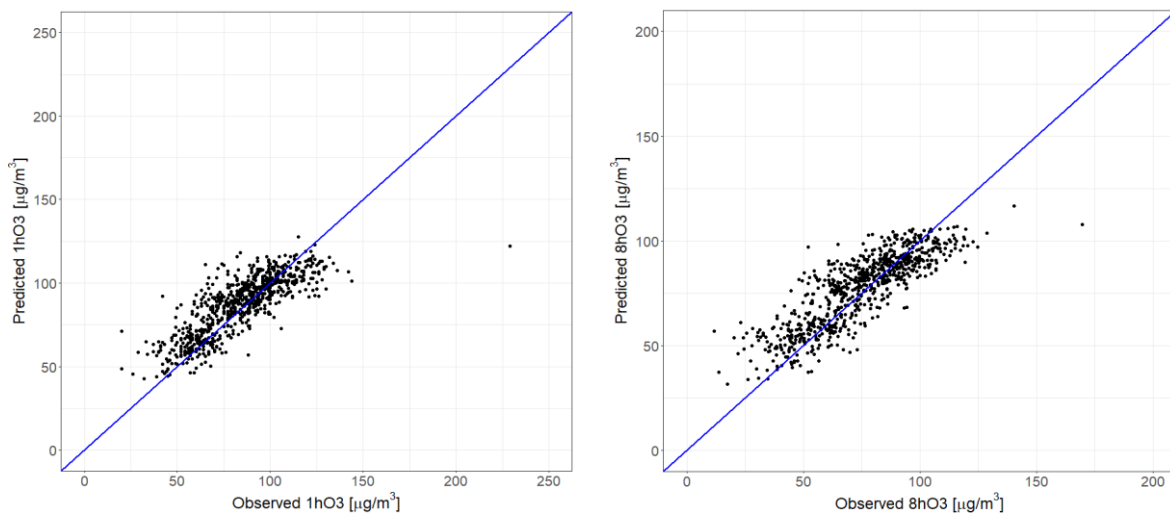


Figure 31 Scattered plot Observed vs Predicted (1hO₃ and 8hO₃) of the whole year for Barcelona – Palau Reial and Barcelona–Zona Universitaria with the testing set

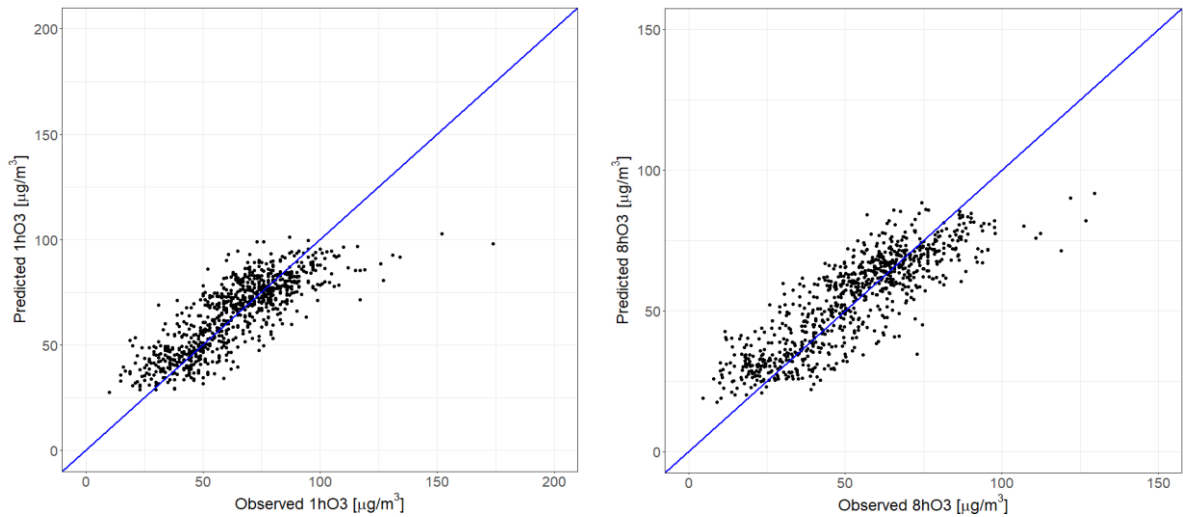


Figure 32 Scattered plot Observed vs Predicted (1hO₃ and 8hO₃) of the whole year for Barcelona–Eixample and Barcelona–El Raval with the testing set

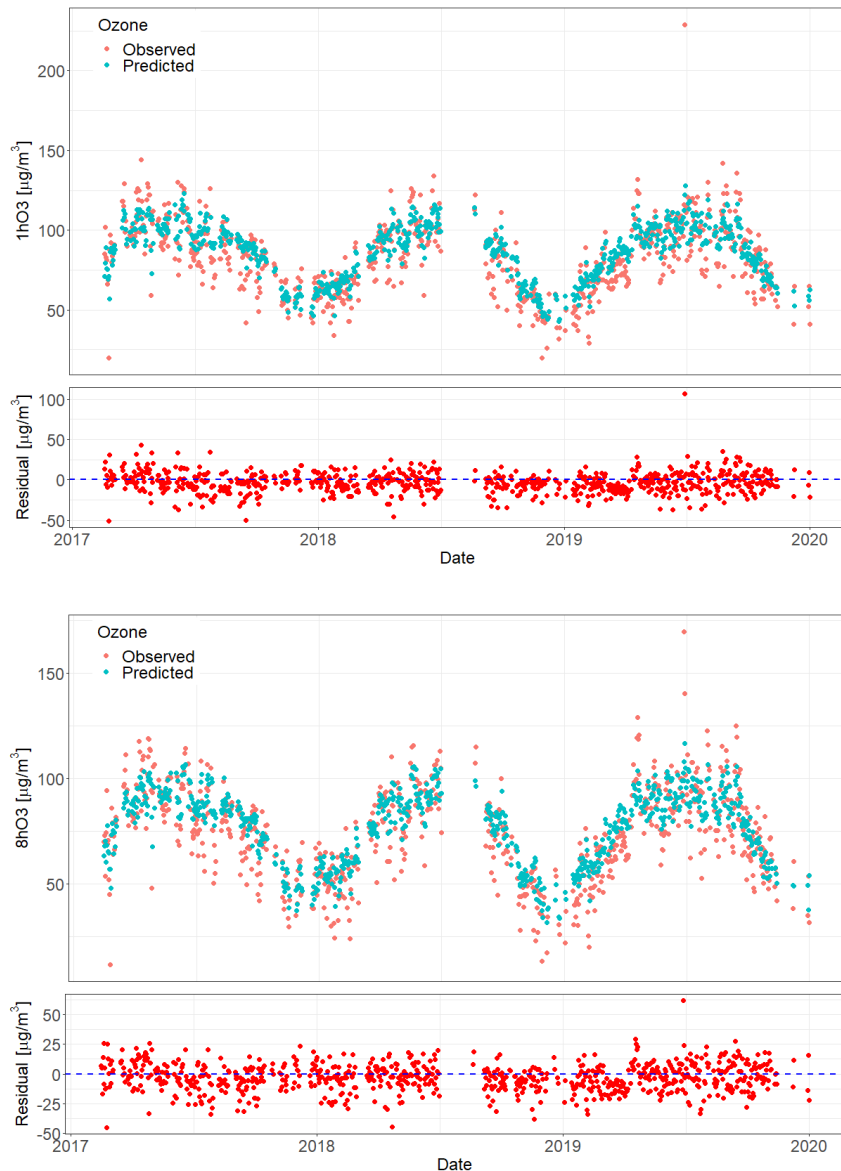


Figure 33 Time series of 1hO₃ and 8hO₃ for the whole year in PR_ZU with the testing set

The predicted outputs, both 1hO₃ and 8hO₃ capture the temporal variation of the ground-level ozone as we can see in Figure 33 and Figure 34, where only the testing period is considered. The models have difficulties to interpret the extraordinarily high or low values.

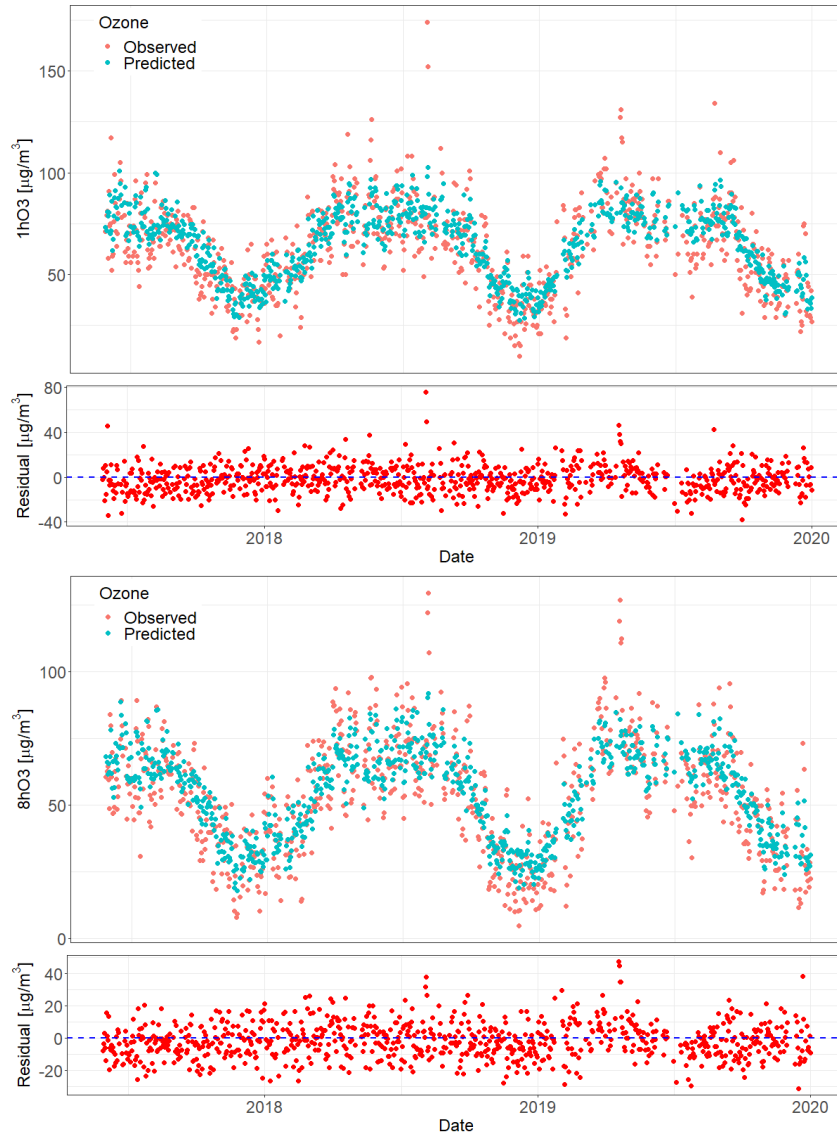


Figure 34 Time series of 1hO₃ and 8hO₃ for the whole year in EI_RA with the testing set

5.1.1.2 Model interpretation

It was explained previously that one of the inputs of the models is the daily maximum hourly ozone concentration level or the daily maximum 8-hours average ozone concentration level of the day before the prediction (O₃). In other words, 1hO₃ or 8hO₃ of the previous day are been used if we are predicting 1hO₃ or 8hO₃ respectively. This variable is the most important one in every regression model considering data of the whole year as we can see in Figure 35 and Figure 36. In order to measure the Variable importance, we have used one of the options included in *randomForest* package, mean decrease impurity (MDI). In the case of regression models, MDI is measured based on the decrease of residual sum of squares (RSS).

The higher the MDI, the higher the importance is. Solar radiation (Solar Rad) and the day of the year (Day_year) complete the podium in the second and third place in importance respectively. Months and the moving average on NO in seven days (NO_7) are also present in the four models as the most important ones. NO acquires higher importance while the period of the moving average is higher.

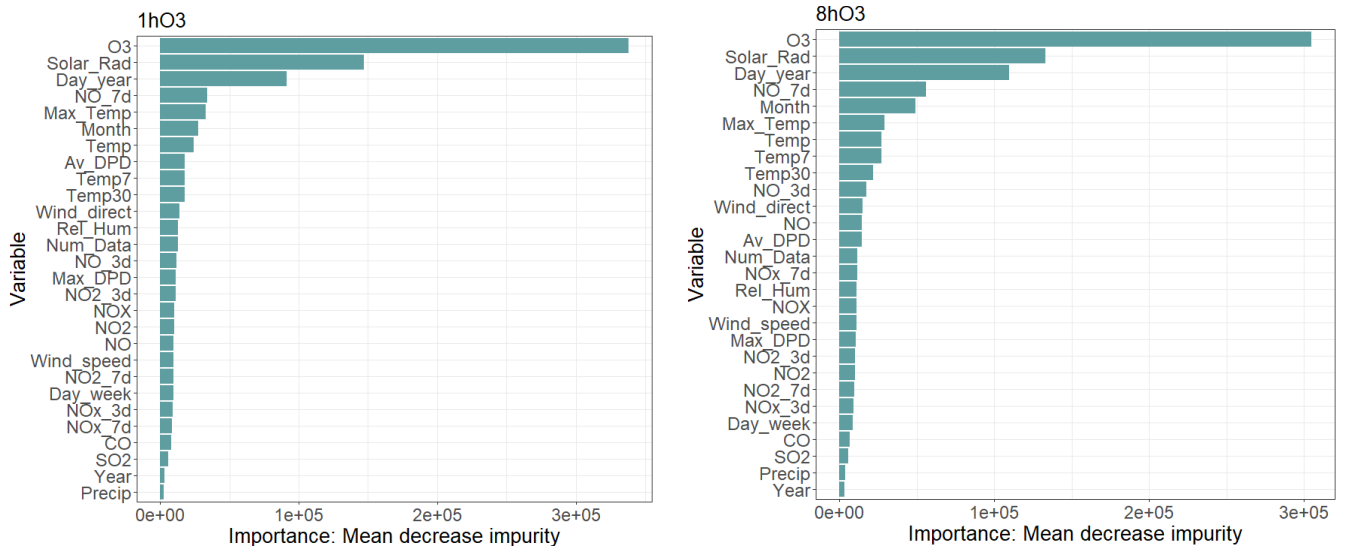


Figure 35 Variable importance for 1hO₃ and 8hO₃ in RF regression models of the whole year in PR_ZU

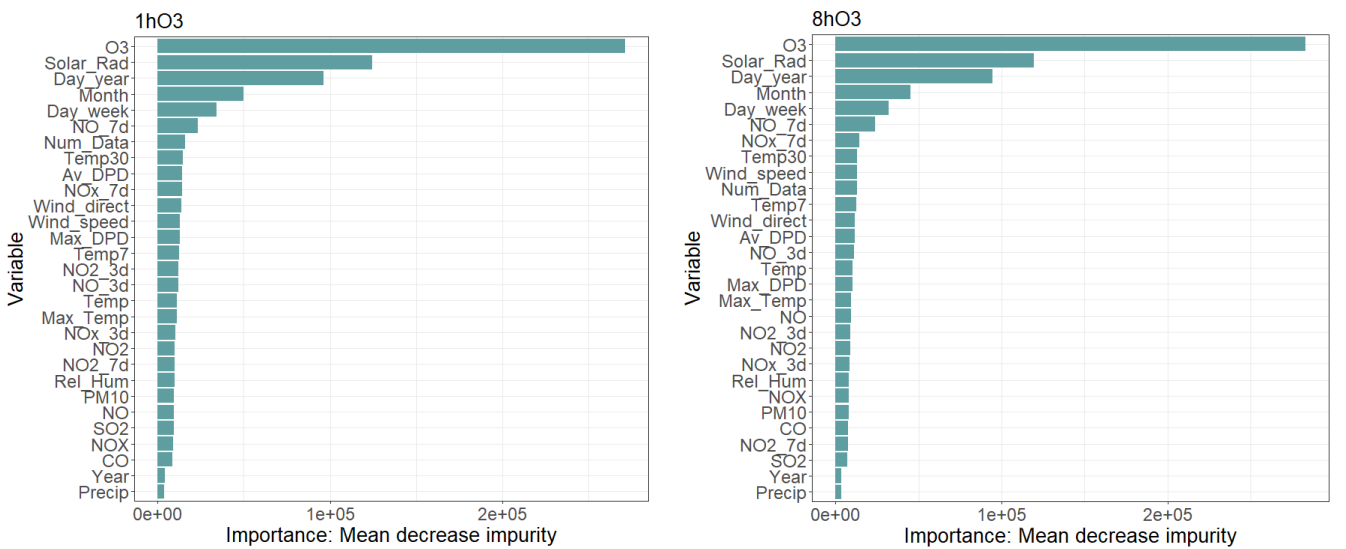


Figure 36 Variable importance for 1hO₃ and 8hO₃ in RF regression models of the whole year in EI_RA

We took the case of 1hO₃ for PR_ZU as an example to see the behaviour of the main variables (Figure 37). Tropospheric ozone of the day before the prediction, solar radiation, and maximum temperature have a positive relationship with the output. The inverse relationship between NO_7 and 1hO₃ might be related to what we studied previously in the exploration step (section 4.2.2). Finally, the model correctly captures the seasonal behaviour as we can see in the variation of 1hO₃ respect to months and days of the years with the highest values in summer.

The highest variation of 1hO₃ takes place with O₃ as a variable because this is the most important one.

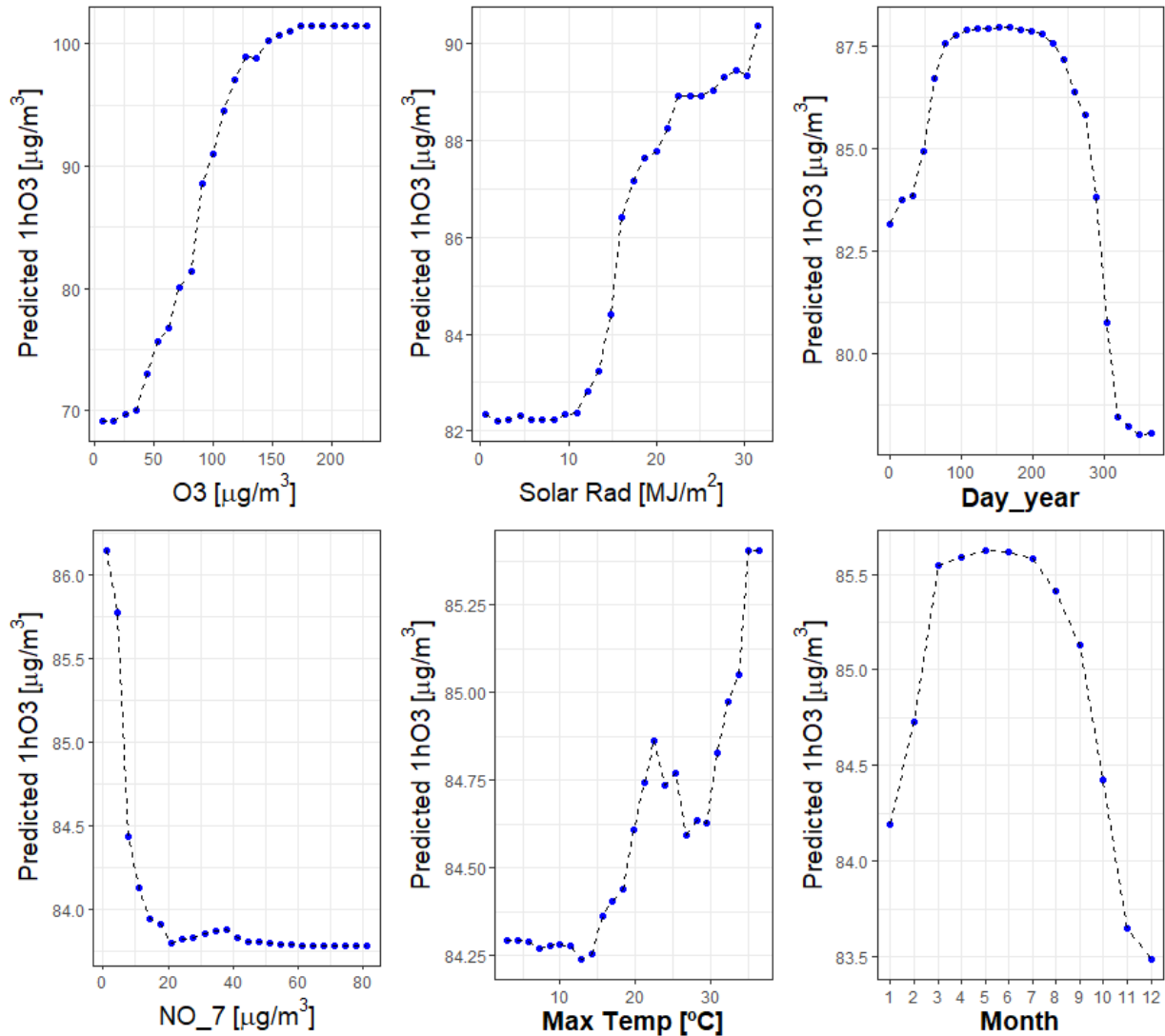


Figure 37 Partial importance of the main variables for 1hO₃ output of whole year in PR_ZU

The generation of ground-level ozone is linked to the presence of NO_x and similar components as we saw previously. Consequently, how 1hO₃ varies with respect to these variables in the model is something important to analyse. In Figure 38, we can see partial importance plots of NO, NO₂, NO_x and their moving averages for PR_ZU. We appreciate a negative correlation between NO and 1hO₃, a positive one between NO₂ and 1hO₃, and no specific correlation between NO_x and 1hO₃. We must understand that NO_x refers to the nitrogen oxides present in the air in a generic form, and its measurement is mainly the combination of NO and NO₂ (Akimoto *et al.*, 2006). The increase in tropospheric ozone concentrations during weekends might be related specifically to NO and its reduction during these days, and even being a reaction produced in presence of sunlight and during several hour of the day (Akimoto *et al.*, 2006), the daily moving averages of NO acquire a higher importance in every model than the maximum value of NO during the day.

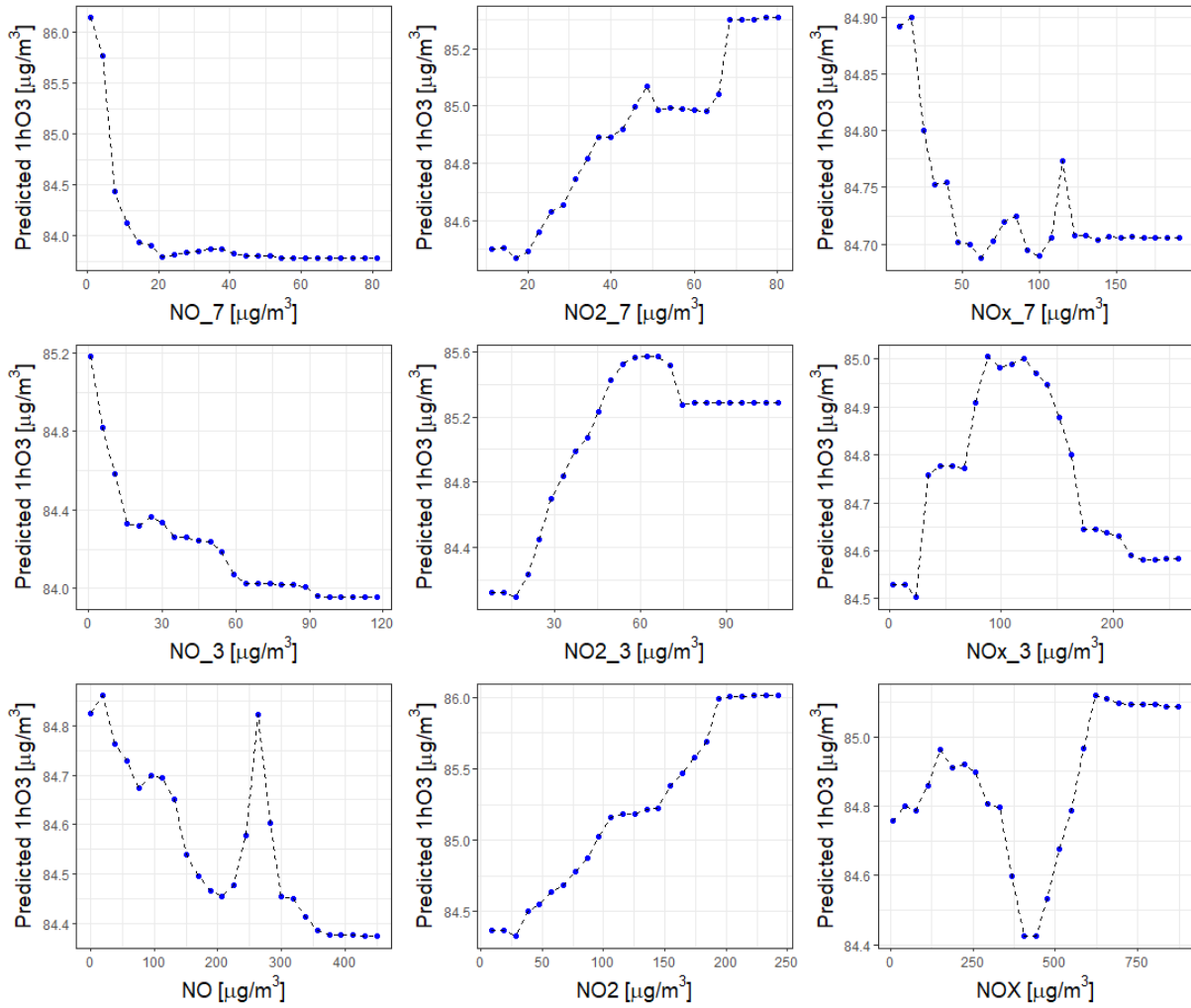


Figure 38 Partial importance of NO, NO₂, NO_x and their moving averages for 1hO₃ output of whole year in PR_ZU

We could see until this point that RF models of the whole year for both outputs have similar behaviour and similar error metrics. Therefore, in order to make some further analysis, we decided to study the behaviour of the models when we remove O₃ as a variable (1hO₃ or 8hO₃ of the previous day) because this is the most important variable by far in every model, and this aspect could mask the potential of the other variables or maybe modify the order of importance. Thus, we took again the model of 1hO₃ for PR_ZU to make this analysis as an example.

	ME	RMSE	MAE	MPE	MAPE
All variables	-3.248	13.423	9.970	-7.195	13.725
Without O ₃	-5.077	15.512	11.812	-10.049	16.409
%Variation	56.30%	15.56%	18.48%	39.66%	19.56%

Table 14 Error metrics for 1hO₃ model with and without O₃ as variable considering the testing set for the whole year in PR_ZU

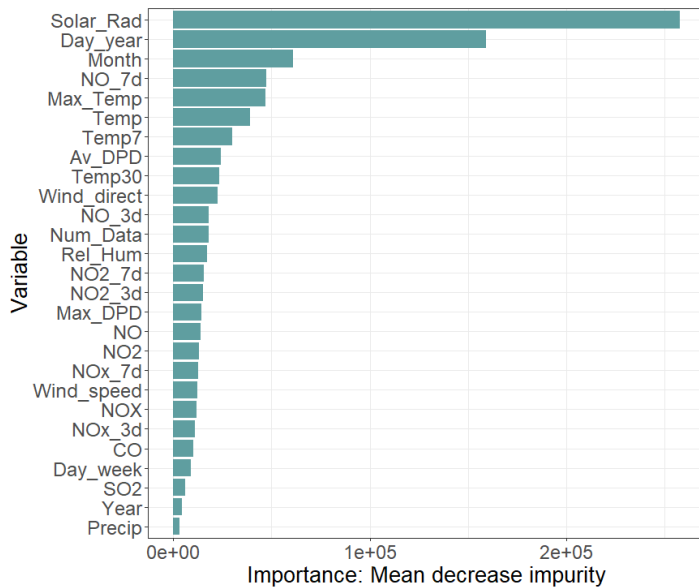


Figure 39 Variable importance for 1hO₃ in RF model of the whole year without O₃ in PR_ZU

Solar radiation and the day of the year are still the two main variables (if we remove ozone level from the previous day) as we can see in Figure 39. However, some variables increase their importance such as months or Temp7. In general terms, the results about importance without O₃ are similar to those obtained with all the variables included. Although, the error metrics after removing O₃ increase considerably as Table 14 shows. This confirms the O₃ importance and showing that it does not influence over the importance of other variables.

In order to show graphically the variation of 1hO₃ with respect to the main two variables, we took the case of PR_ZU as an example, Figure 40 shows that while O₃ and Solar Rad grow, 1hO₃ also grows. The combination of both variables can deliver a wide 1hO₃ variation.

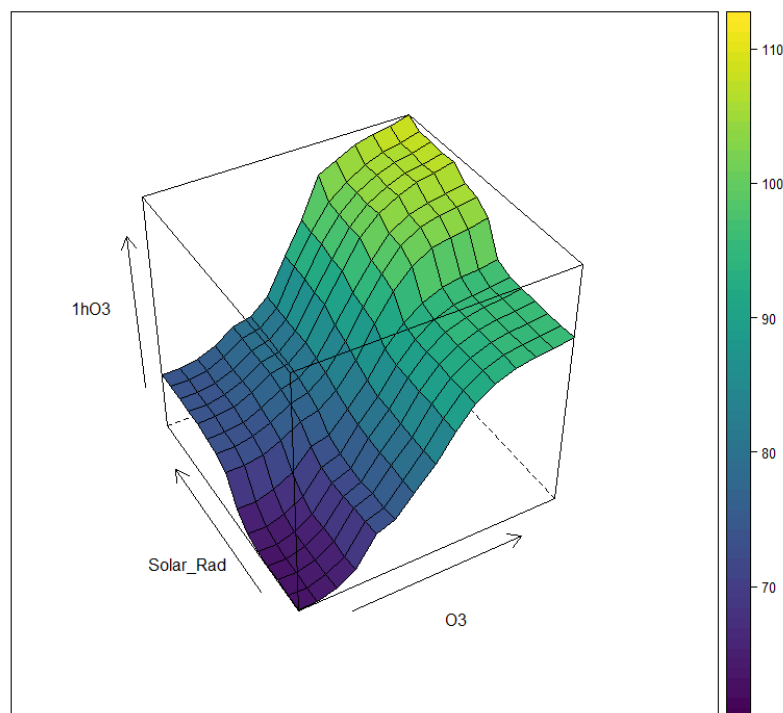


Figure 40 1hO₃ variation with respect to Solar Rad and O₃ for PR_ZU using data of the whole year

5.1.2 Categorical model

In section 2.11, we saw that the way to evaluate the accuracy of a categorical model (a model with a categorical output) is calculating the confusion matrix and through it, the error rate of the model. The results of the average error rate from the application of the model (with different parameters) over the testing set of the folds in the prequential evaluation analysis for 1hO₃ in PR_ZU is shown in Figure 41. This is an example of the procedure to select the parameters, which produce the highest accuracy. We can see that 800 trees (*ntrees*) and 14 features or variables (*mtry*) are the most accurate combination with an error rate of 0.1356. The procedure is repeated for the categorical outputs of 1hO₃ and 8hO₃ in both couples of stations and the parameters with the lowest error rate are summarized in Table 15.

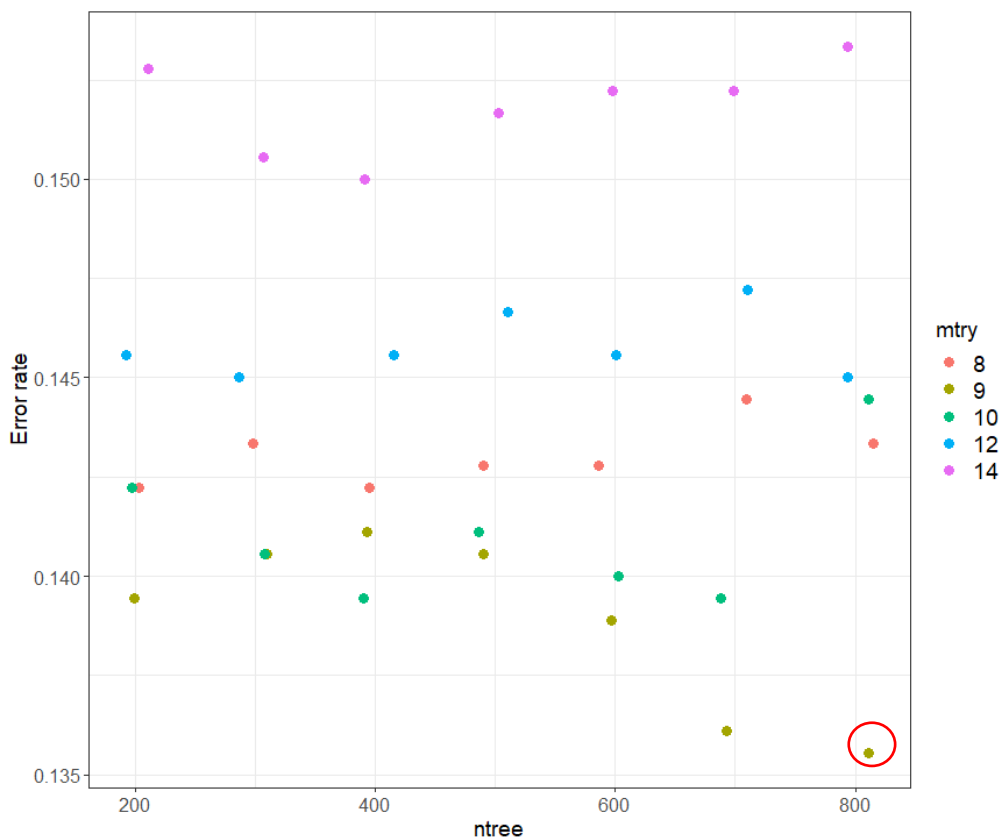


Figure 41 Mean error rate of the prequential evaluation analysis for categorical model of 1hO₃ in PR_ZU with data of the whole year using several combinations of *ntree* and *mtry*

Output	Station	<i>ntree</i>	<i>mtry</i>	Error rate
1hO ₃	PR_ZU	800	9	0.1356
8hO ₃	PR_ZU	400	8	0.3906
1hO ₃	EI_RA	200	14	0.1267
8hO ₃	EI_RA	300	8	0.3478

Table 15 Summary of the results of the prequential evaluation for the most accurate combination of parameters for every categorical model of the year

Once we determined the best parameters for the models, we trained a new model using the whole training set, and the resulting model was applied to predict the categories of 1hO₃ and 8hO₃ using the variables of the testing set. As a result of this procedure, we got the out-of-bag (OOB) and testing error rates. We saw in section 2.6, that OOB analysis is also a good accuracy measure for the model.

		Confusion Matrix	1hO ₃ < 86	1hO ₃ ≥ 86	Error rate
			Error rate of class 1	Error rate of class 2	
Barcelona-Palau Reial and Barcelona-Zona Universitaria	OOB	$\begin{vmatrix} 792 & 147 \\ 114 & 752 \end{vmatrix}$	0.1565	0.1316	0.1446
	Testing set	$\begin{vmatrix} 284 & 111 \\ 38 & 341 \end{vmatrix}$	0.2810	0.1003	0.1925
Barcelona-Eixample and Barcelona-El Raval	OOB	$\begin{vmatrix} 1641 & 47 \\ 177 & 91 \end{vmatrix}$	0.0278	0.6604	0.1145
	Testing set	$\begin{vmatrix} 681 & 41 \\ 66 & 50 \end{vmatrix}$	0.0568	0.5689	0.1277

Table 16 Error rate of out-of-bag samples and testing set for categorical 1hO₃ in PR_ZU and EI_RA for the model of the whole year

Confusion matrix of the OOB analysis and testing set along with the error rate for the whole model and each category are shown in Table 16 and Table 17. In every case the OOB error rate is closer to the one obtained in the prequential evaluation analysis than the error rate gotten with the testing set. However, their magnitude is similar. The error rate of the whole categorical model of 1hO₃ is relatively low in both couples of stations. Nonetheless, the error is considerable in the categorical models of 8hO₃.

In PR_ZU, both categories for 1hO₃ show a low error. However, this is not the case in EI_RA, as we saw when we explored the data, there are few values inside the second category selected ($\geq 86 \mu\text{g}/\text{m}^3$); therefore, it is difficult to train the model and obtain lower error rates. The same principle is applied to the 3rd and 4th categories of 8hO₃ for EI_RA where the error rate is high (0.73 and 0.65 for the 3rd, and 0.95 and 0.89 for the 4th). The first category for both 1hO₃ and 8hO₃ ($1\text{hO}_3 < 86 \mu\text{g}/\text{m}^3$ and $0 \leq 8\text{hO}_3 < 55$ respectively) generally presents an error rate lower than 0.25, except for the one in the testing set for 8hO₃ in PR_ZU where the error is considerably high.

The results show the importance to have a uniform distribution for the categories. Nevertheless, this is very complex because we should have a different category scale for every station according to the values of $1hO_3$ and $8hO_3$, defining what is good or what is an unhealthy level in every case. Therefore, although there can be a low overall error rate of the model, the model might have a low accuracy predicting values of specific categories.

		Confusion Matrix	$0 \leq 8hO_3 < 55$	$55 \leq 8hO_3 < 71$	$71 \leq 8hO_3 < 86$	$8hO_3 \geq 86$	Error rate
			Error rate of class 1	Error rate of class 2	Error rate of class 3	Error rate of class 4	
Barcelona-Palau Reial and Barcelona-Zona Universitaria	OBB	$\begin{vmatrix} 312 & 84 & 10 & 0 \\ 112 & 150 & 82 & 26 \\ 21 & 77 & 166 & 146 \\ 2 & 11 & 107 & 495 \end{vmatrix}$	0.2315	0.5945	0.5951	0.1951	0.3765
	Testing set	$\begin{vmatrix} 85 & 59 & 8 & 4 \\ 17 & 70 & 58 & 21 \\ 2 & 24 & 77 & 100 \\ 0 & 2 & 33 & 212 \end{vmatrix}$	0.4551	0.5783	0.6207	0.1417	0.4249
Barcelona-Eixample and Barcelona-El Raval	OBB	$\begin{vmatrix} 863 & 164 & 4 & 0 \\ 157 & 381 & 51 & 1 \\ 26 & 164 & 71 & 6 \\ 0 & 19 & 44 & 3 \end{vmatrix}$	0.1629	0.3542	0.7341	0.9545	0.3254
	Testing set	$\begin{vmatrix} 342 & 100 & 1 & 0 \\ 35 & 176 & 31 & 1 \\ 5 & 62 & 37 & 2 \\ 0 & 12 & 29 & 5 \end{vmatrix}$	0.2279	0.2757	0.6509	0.8913	0.3317

Table 17 Error rate of out-of-bag samples and testing set for categorical $8hO_3$ in PR_ZU and EI_RA for the model of the whole year

MDI in categorical models is measured based on Gini index; therefore, this can also be called mean decrease Gini (MDG). We obtained the Variable importance for every model based on MDG (Figure 42 and Figure 43). O_3 , solar radiation and day of the year are the main variables for the categorical models (as we had previously for the regression ones) in both couples of stations. However, there are some differences between results, in EI_RA, weekday (Day_week) acquires a notorious importance along with the number of the data or sample (Num_Data). The moving average of NO in seven days (NO_7) is once again the most important air quality variable in almost all models, except for NO2_3 in EI_RA categorical $1hO_3$. The year and precipitation have the lowest importance.

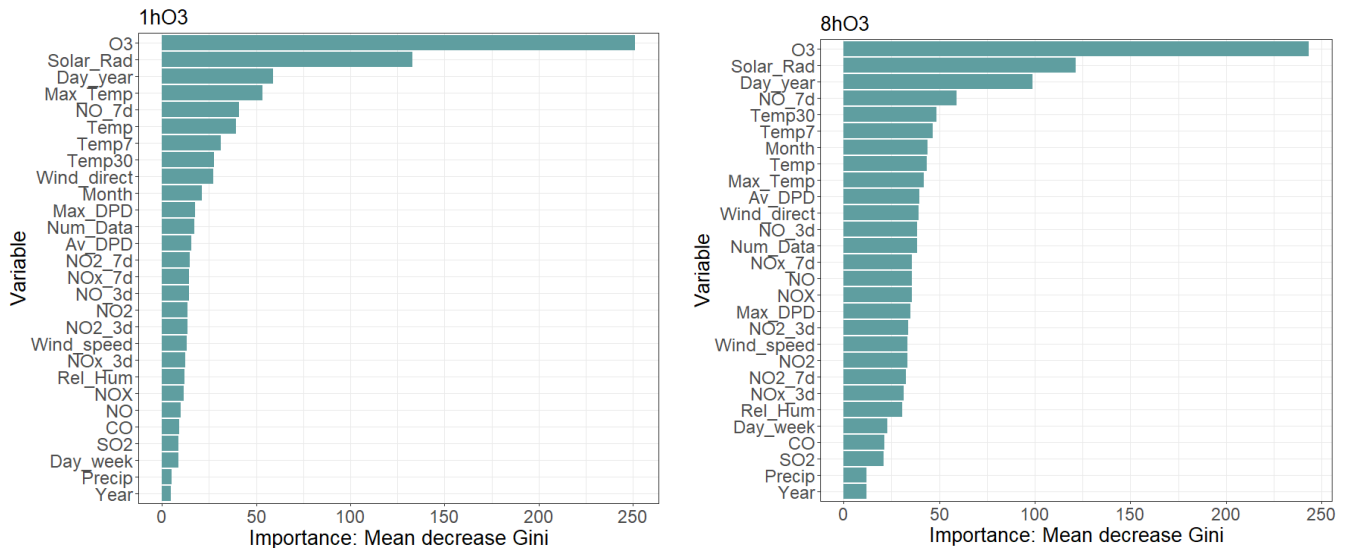


Figure 42 Variable importance for 1hO₃ and 8hO₃ in RF categorical models of the whole year in PR_ZU

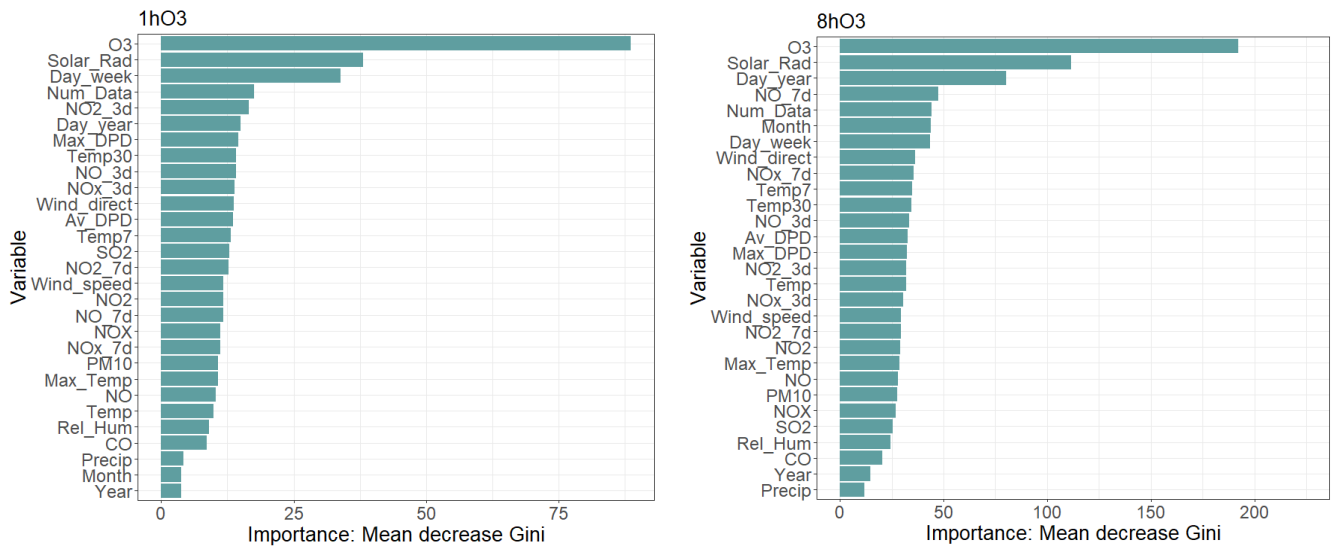


Figure 43 Variable importance for 1hO₃ and 8hO₃ in RF categorical models of the whole year in EI_RA

Following the same analysis that we did in the previous section; we considered the same set of inputs without ground-level ozone (O₃) to see the performance and behaviour of the model and as an example we took Barcelona-Palau Reial and Barcelona-Zona Universitaria. About the categories of 1hO₃, there is not a considerable difference in error rates between Table 16 and Table 18 when we do not take into account O₃ as input, neither in the general error rate of the model nor in the categories; although, the highest variation is appreciated in the lowest class.

Barcelona-Palau Reial and Barcelona-Zona Universitaria		Confusion Matrix	1hO ₃ < 86	1hO ₃ ≥ 86	Error rate
			Error rate of class 1	Error rate of class 2	
	OOB	$\begin{vmatrix} 777 & 162 \\ 102 & 764 \end{vmatrix}$	0.1725	0.1177	0.1463
	Testing set	$\begin{vmatrix} 268 & 127 \\ 38 & 341 \end{vmatrix}$	0.3215	0.1003	0.2132

Table 18 Error rate of out-of-bag samples and testing set for categorical 1hO₃ in PR_ZU for the model of the whole year without O₃

In the analysis with categories of 8hO₃ (Table 17 and Table 19), the variation of OOB error rate is low; however, in the testing set, the error rate is higher in lower classes and in the last category, the error rate is lower than we saw previously (complete model). Thus, solar radiation (the most important variable in this case) might be crucial especially for prediction of high levels of tropospheric ozone. In general terms, error rate is higher when we do not take into account O₃, and the order of the variation is practically the same (Figure 44).

Barcelona-Palau Reial and Barcelona- Zona Universitaria		Confusion Matrix	0 ≤ 8hO ₃ < 55	55 ≤ 8hO ₃ < 71	71 ≤ 8hO ₃ < 86	8hO ₃ ≥ 86	Error rate
			Error rate of class 1	Error rate of class 2	Error rate of class 3	Error rate of class 4	
	OOB	$\begin{vmatrix} 316 & 74 & 16 & 0 \\ 109 & 148 & 72 & 41 \\ 28 & 77 & 131 & 174 \\ 1 & 10 & 85 & 519 \end{vmatrix}$	0.2217	0.6000	0.6805	0.1561	0.3815
	Testing set	$\begin{vmatrix} 72 & 58 & 22 & 4 \\ 19 & 52 & 46 & 49 \\ 1 & 22 & 56 & 124 \\ 0 & 2 & 20 & 225 \end{vmatrix}$	0.5384	0.6867	0.7241	0.0891	0.4754

Table 19 Error rate of out-of-bag samples and testing set for categorical 8hO₃ in PR_ZU for the model of the whole year without O₃

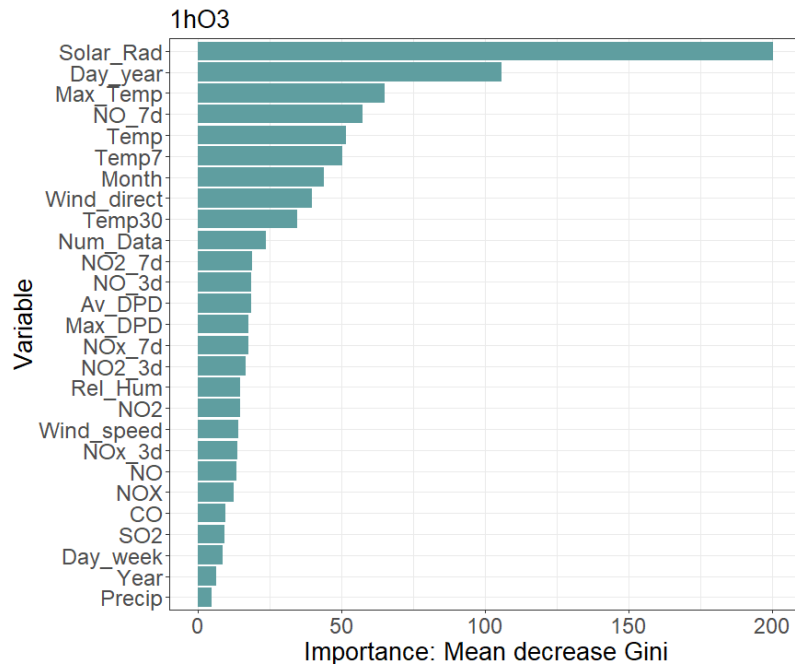


Figure 44 Variable importance for 1hO₃ in RF categorical model of the whole year without O₃ in PR_ZU

5.2 Models from May to September

5.2.1 Regression models

5.2.1.1 Prediction accuracy

The addition of the J&C synoptic classification (SC) is the main characteristic of these models as we saw previously. This variable is added to the models as a categorical input, the only one that our models have. The prequential evaluation results for 1hO₃ for PR_ZU with days from May to September can be seen in Figure 45. In this example, we obtained that the combination of parameters with the lowest MAE is $n_{tree} = 200$ and $m_{try} = 14$. The results of the most accurate combinations of parameters for every regression model in both couples of stations are shown in Table 20. Average MAE varies between 8.35 and 10.12, which is consistent with the MAE results obtained with the regression models of the whole year (Table 11). 8hO₃ models have the lowest MAE, this is related with lower values compared to 1hO₃.

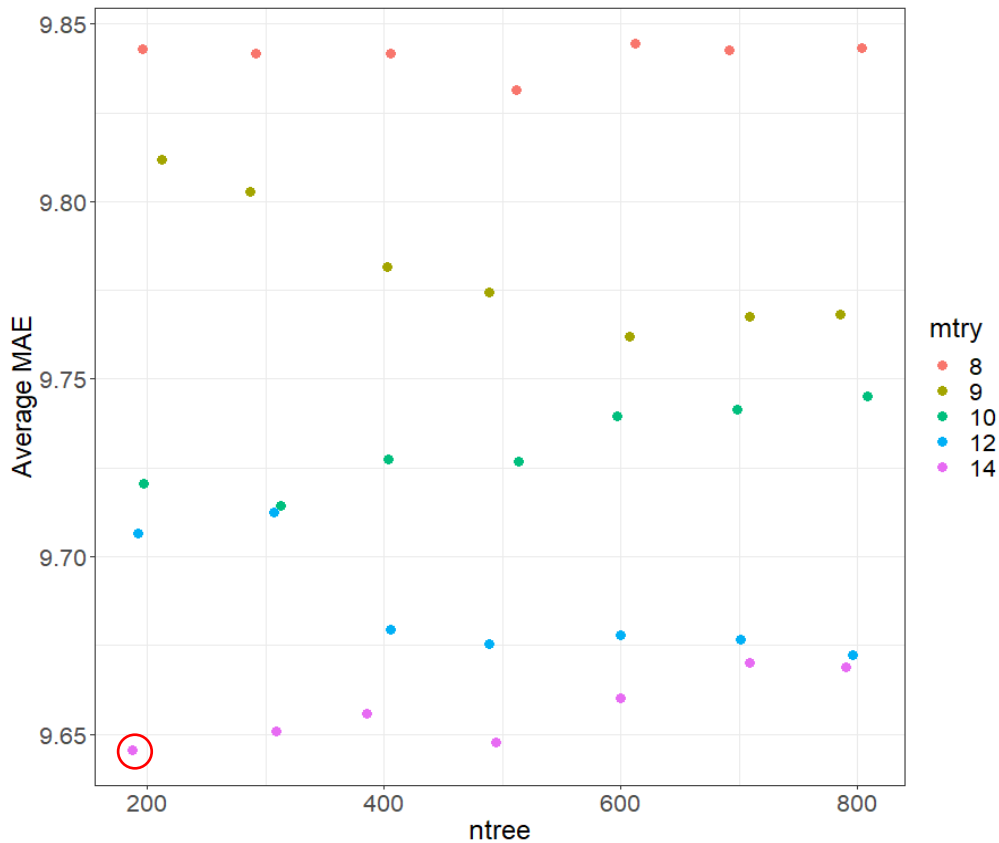


Figure 45 Average MAE of the prequential evaluation analysis for 1hO3 in PR_ZU with data from May to September using several combinations of *ntree* and *mtry*

Output	Station	<i>ntree</i>	<i>mtry</i>	MAE
1hO ₃	PR_ZU	200	14	9.646
8hO ₃	PR_ZU	600	14	8.354
1hO ₃	EI_RA	800	14	10.115
8hO ₃	EI_RA	300	14	8.980

Table 20 Summary of the results of the prequential evaluation for the most accurate combination of parameters for every regression model for data from May to September

Once we selected the parameters for our models, we applied them to build a new model using the whole training set. Afterward, the inputs of the testing set were introduced to the trained model and the predicted values were compared with the observed ones. The results of the error metrics can be appreciated in Table 21. MAE values are close to the ones obtained in the prequential evaluation. MAPE between observed and predicted varies between 12.3% and 14.7%. These results are lower than the ones that we obtained in the models of the whole year. This means that on average the predicted values have a small percentage difference from the observed ones. In Table 22, we have RMSE after the analysis OOB in the training process. The results are similar to the ones obtained in the testing set.

		ME	RMSE	MAE	MPE	MAPE
PR_ZU	1hO ₃	-3.603	15.084	10.928	-6.435	12.569
	8hO ₃	-2.922	12.307	9.606	-5.821	12.291
EI_RA	1hO ₃	-1.799	13.642	10.373	-5.714	14.727
	8hO ₃	-0.976	10.931	8.576	-4.657	14.111

Table 21 Error metrics of the testing set of the RF models with data from May to September for every output in PR_ZU and EI_RA

		RMSE.OOB
PR_ZU	1hO ₃	12.493
	8hO ₃	10.533
EI_RA	1hO ₃	12.394
	8hO ₃	10.889

Table 22 Out-of-bag RMSE for 1hO₃ and 8hO₃ with data from May to September

The scattered plots of the observed and predicted values made with the testing set (Figure 46 and Figure 47) for both couples of stations do not follow consistently the ideal tendency shown with blue line. This indicates that the model has difficulties capturing the temporal variation of ground-level ozone when we only consider days from May to September, this situation did not occur with a model of the whole year. EI_RA shows a higher dispersion of the values than PR_ZU, especially with the model of 8hO₃.

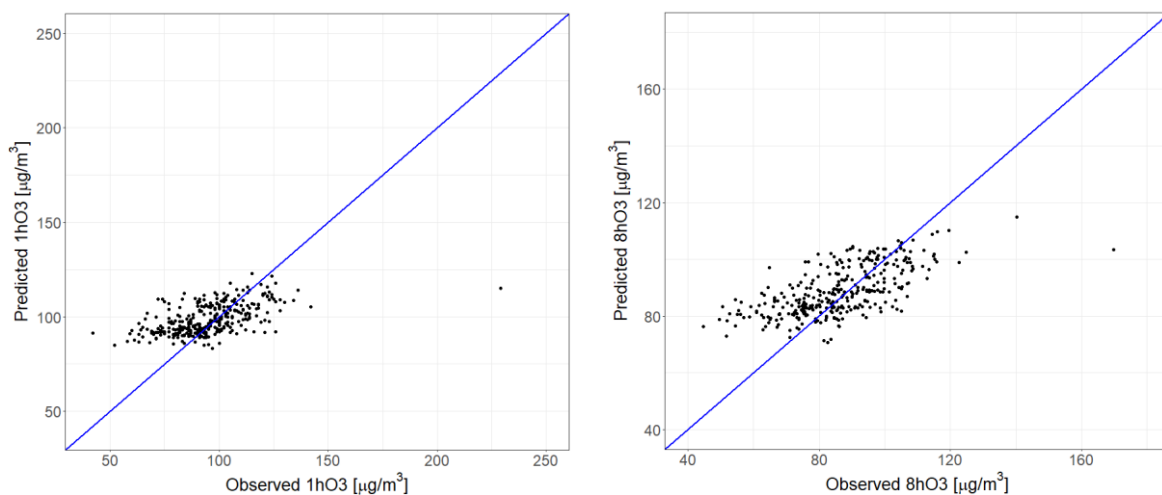


Figure 46 Scattered plot of Observed vs Predicted (1hO₃ and 8hO₃) of days from May to September for Barcelona – Palau Reial and Barcelona–Zona Universitaria with the testing set

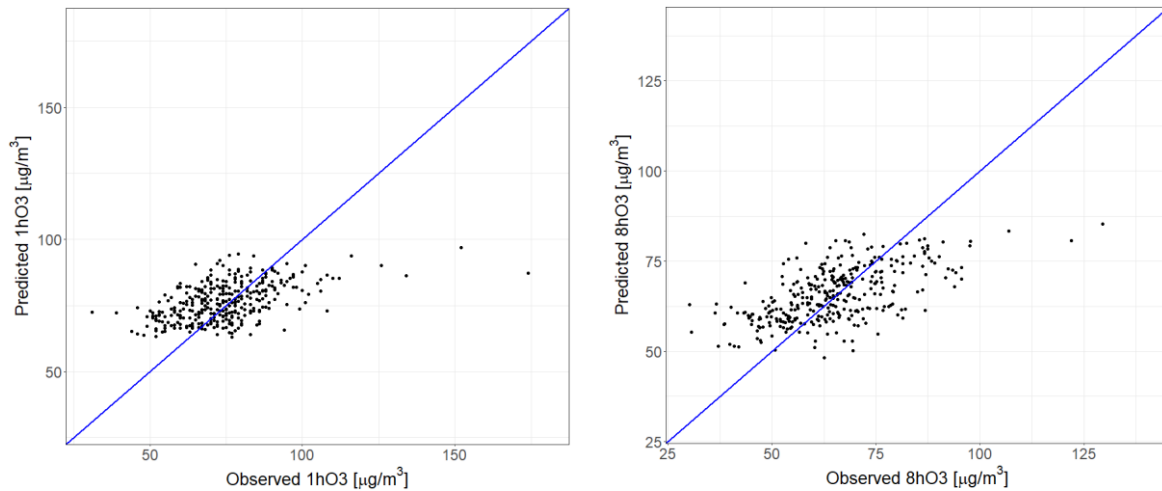


Figure 47 Scattered plot of Observed vs Predicted (1hO₃ and 8hO₃) of days from May to September for Barcelona – Eixample and Barcelona–El Raval with the testing set

5.2.1.2 Model interpretation

Special attention must be paid to the determination of variable importance in RF models with categorical inputs, since the implementation of RF method in *randomForest* package has bias toward categorical variables, especially if this has many categories (Strobl *et al.*, 2007). Therefore, we verified such possible bias by adding two randomly generated categorical variables as inputs —cat1 and cat2— with 10 and 40 categories respectively, and computing the variable importance in *randomForest* and *party* (with *cforest* as main function) packages. We used for this verification the model of 1hO₃ in PR_ZU.

We commented before that there are two ways to measure the variable importance, MDI and MDA. Both measurements are shown in Figure 48 when we compute variable importance with *randomForest* function adding the two variables that we mentioned before. Both cat1 and cat2 have a high importance when we consider MDI, cat2 importance is even close to O₃, followed by SC. This situation is not correct because cat2 and cat1 are variables that we just made up, which do not have any connection with the output and cannot have this importance. This situation changes when we consider MDA; however, cat2 still have some importance. The results of this simple verification make us think that *randomForest* can deliver unreliable variable importance when we consider categorical inputs.

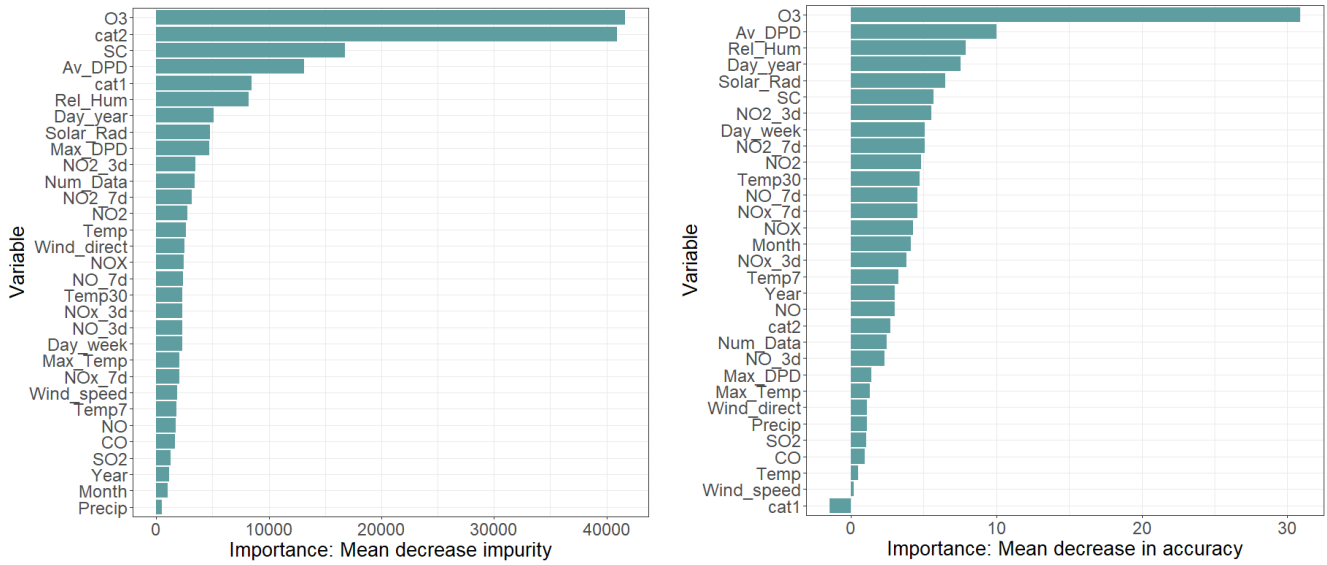


Figure 48 Variable importance for 1hO₃ in RF regression model in PR_ZU using *randomForest* function with data from May to September adding two randomly generated categorical variables

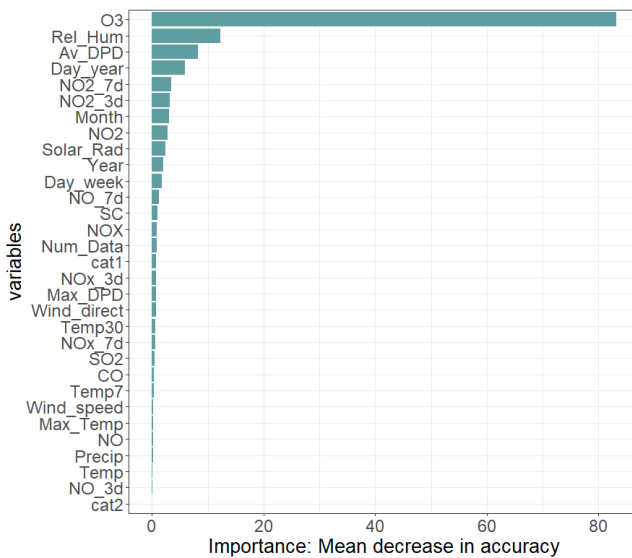


Figure 49 Variable importance for 1hO₃ in RF regression model in PR_ZU using *cforest* with data from May to September adding two randomly generated categorical variables

When we compute variable importance using *cforest* function, we have a different result (Figure 49). Even using MDA, there is a different approach between how *randomForest* and *cforest* create the random forests in the algorithm (Strobl *et al.*, 2007), this is why the variable importance results are different. In the case of *cforest*, cat2 has meaningless importance and cat1 has a very low importance too, which is something that we would expect. Based on these results and the study of Strobl *et al.*, (2007), we decided to use *cforest* when we add SC to our inputs. Something important to highlight is that the most important numerical inputs are the same in every analysis.

We can also see in Table 23 that the accuracy of both packages is similar taking as an example 1hO₃ in PR_ZU. Hence, the most significant variation of both packages is related with variable importance.

		ME	RMSE	MAE	MPE	MAPE
1hO ₃	<i>randomForest</i>	-4.817	15.617	11.394	-7.893	13.269
	<i>cforest</i>	-3.603	15.084	10.928	-6.435	12.569

Table 23 Comparison of error metrics for the testing set of 1hO₃ models with *randomForest* and *cforest* package for PR_ZU with data from May to September

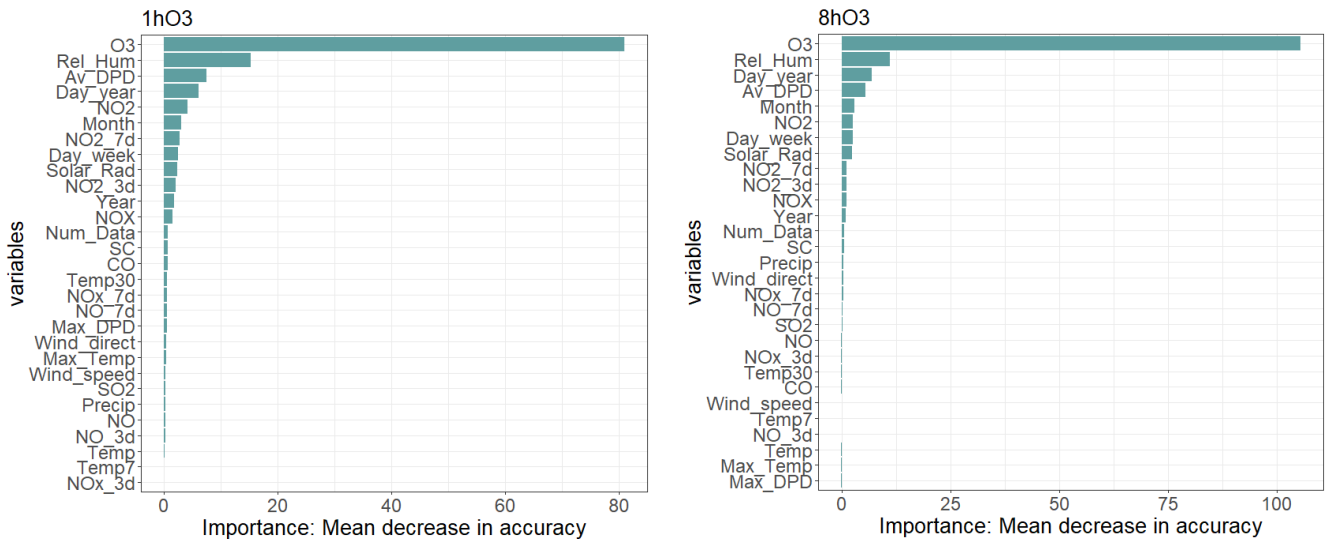


Figure 50 Variable importance for 1hO₃ and 8hO₃ in RF regression models of days from May to September in PR_ZU

O₃, relative humidity, Av DPD, Day of the year, NO₂, and Month are the most important variables in the case of PR_ZU (Figure 50), the main difference with respect to the variable importance of the models of the whole year is that Solar Radiation is not one of the main variables. However, when we took the case of EI_RA (Figure 51), Solar Rad is once again one of the main variables; although, not with the same intensity. This might be related to that Solar Rad does not have a significant variation in summer days; therefore, it provides less information to the model. One of the most important variables is Av DPD confirming that Rel Hum has considerable importance in summer days, especially in regions where contaminants such as NO₂ do not have a high influence (PR_ZU). However, the influence of weekdays is important in EI_RA, this is related to the emission of pollutants in the area as we analysed previously. SC has low importance in every model, especially in EI_RA.

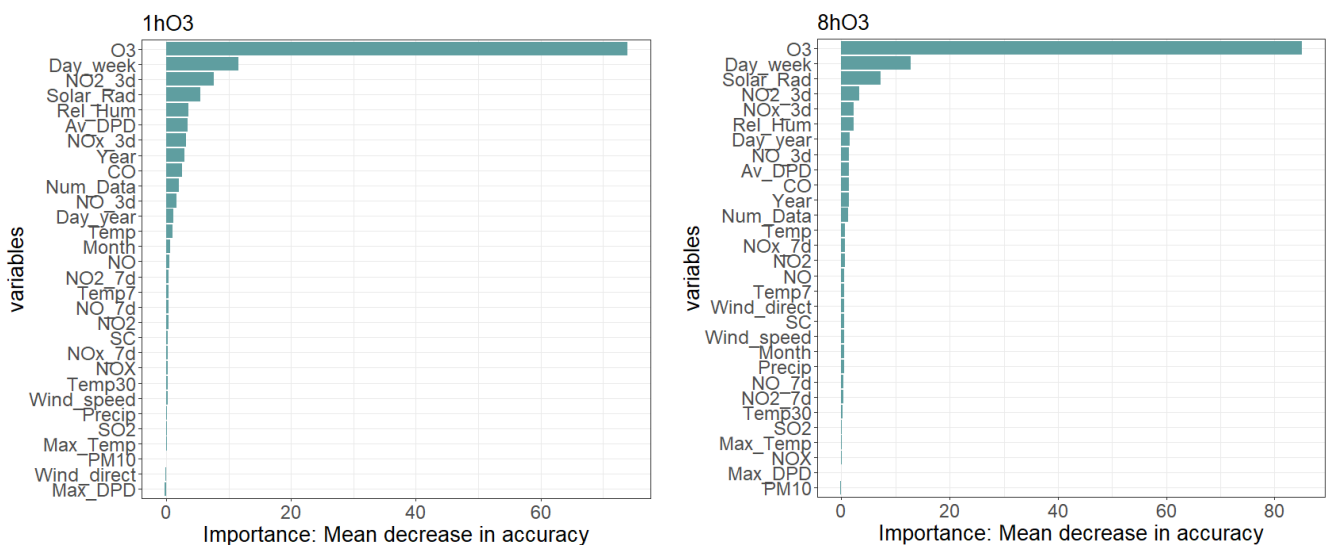


Figure 51 Variable importance for 1hO₃ and 8hO₃ in RF regression models of days from May to September in EI_RA

The partial importance analysis of 1hO₃ in PR_ZU for days from May to September (Figure 52) shows again the positive relationship between O₃ and 1hO₃ as the main variable, the inverse relationship between relative humidity (Rel Hum) and 1hO₃, which is confirmed with Av DPD positively related to 1hO₃ because the higher the Rel Hum the lower DPD will be. In days of September, we have a reduction of tropospheric ozone levels as the relationship between Day_year and Month with 1hO₃ shows. Finally, NO₂ is positively related to 1hO₃, this result is similar to what we obtained in the model of the whole year, the positive correlation indicates that the increase of ozone levels during weekends might not be related to NO₂ levels in these days.

Pure advection and advection with anticyclonic characteristics coming from the east are the atmospheric circulations, which are related to the highest values of tropospheric ozone as we can see in Figure 53, taking PR_ZU as an example.

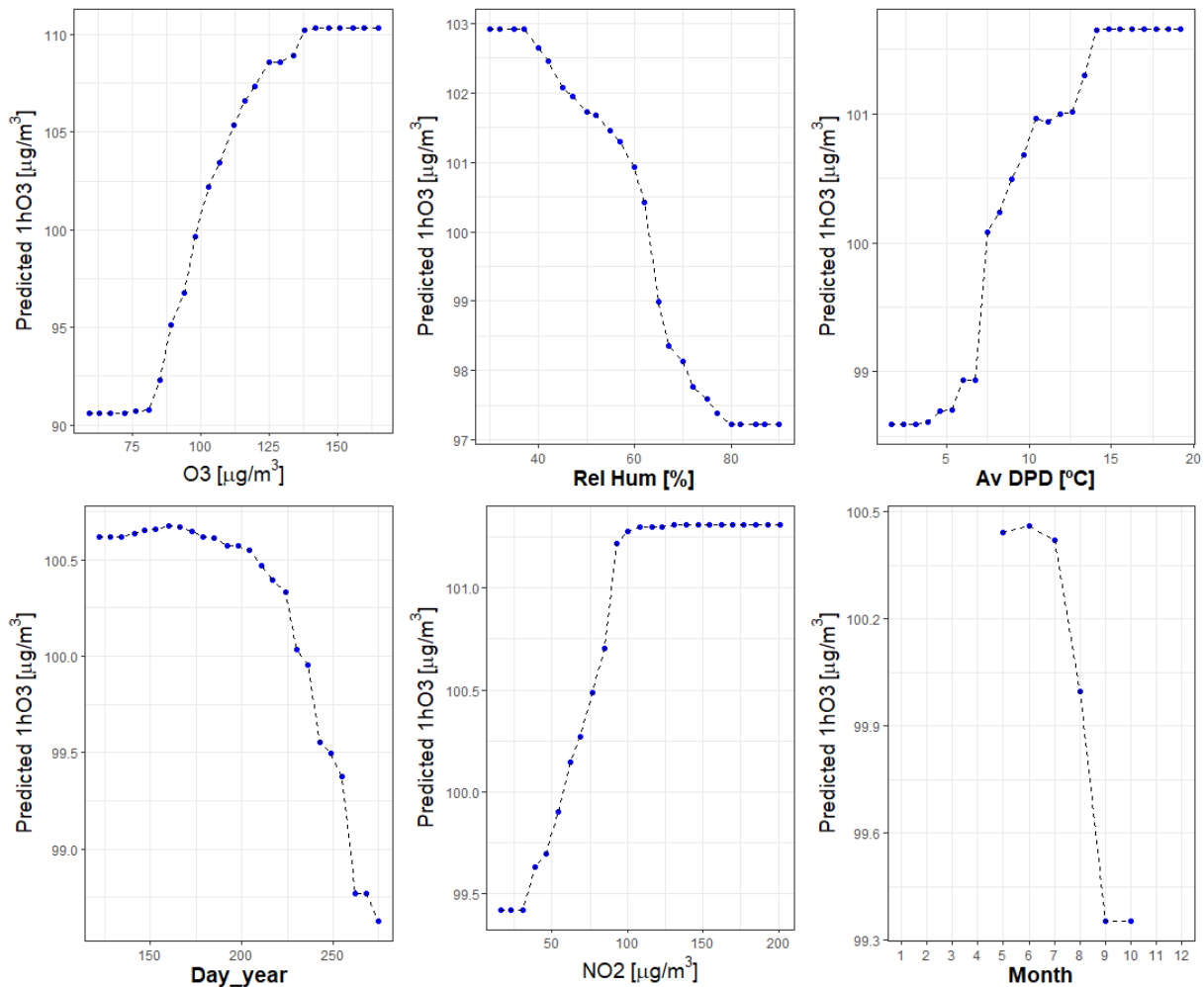


Figure 52 Partial importance of the main variables for 1hO₃ output with data from May to September in PR_ZU

We saw that O₃ has in every case the highest importance. Hence, we do the same analysis, running the models without this variable to see if it is affecting the others masking their potential. In Table 24, we can appreciate that removing O₃ from the variables has an impact increasing every error metrics with respect to the complete model with MAE and MAPE 11% higher.

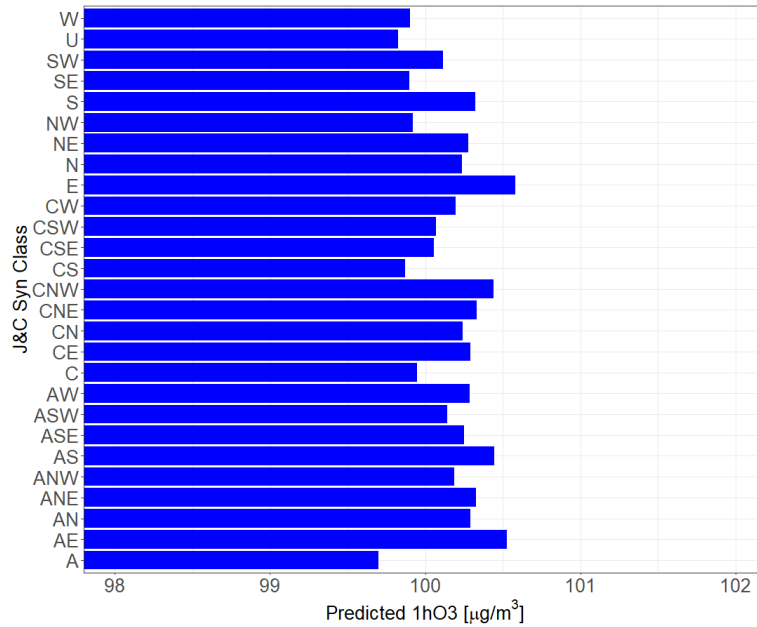


Figure 53 Partial importance of SC for 1hO₃ output with data from May to September in PR_ZU

	ME	RMSE	MAE	MPE	MAPE
All variables	-3.603	15.084	10.928	-6.435	12.569
Without O₃	-4.687	16.432	12.138	-7.857	14.003
%Variation	30.09%	8.94%	11.07%	22.10%	11.41%

Table 24 Error metrics for 1hO₃ model with and without O₃ as variable considering the testing set for days from May to September in PR_ZU

Without O₃, the model (1hO₃ for PR_ZU) preserves Rel Hum and Av DPD as the main variables (Figure 54). In general terms, the order of the variables has not been modified drastically, with the variables obtaining a similar importance to the one that they had in the model with O₃. The presence of NO₂ acquires more importance in summer days. This pollutant and its moving averages are part of the main variables of the model. SC maintains its position not being part of the main variables.

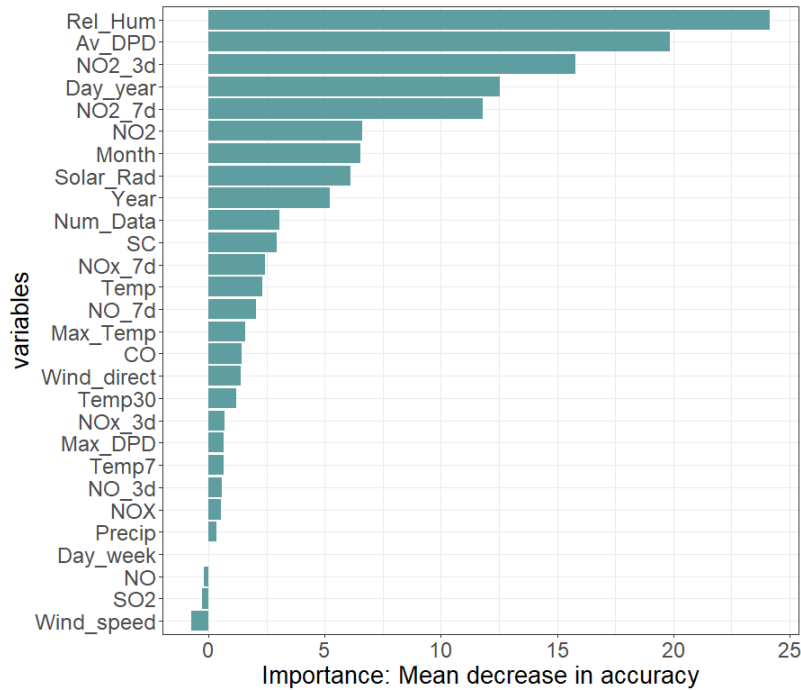


Figure 54 Variable importance for 1hO₃ in RF regression model of days from May to September without O₃ in PR_ZU

Considering again the model with all variables, we can see the variation of 1hO₃ with respect to Rel Hum and O₃ in Figure 55, being both variables the most important ones for 1hO₃ model of days from May to September in PR_ZU. The maximum values of ozone are reached when Rel Hum is low but O₃ is high. A similar result is also shown individually in Figure 52.

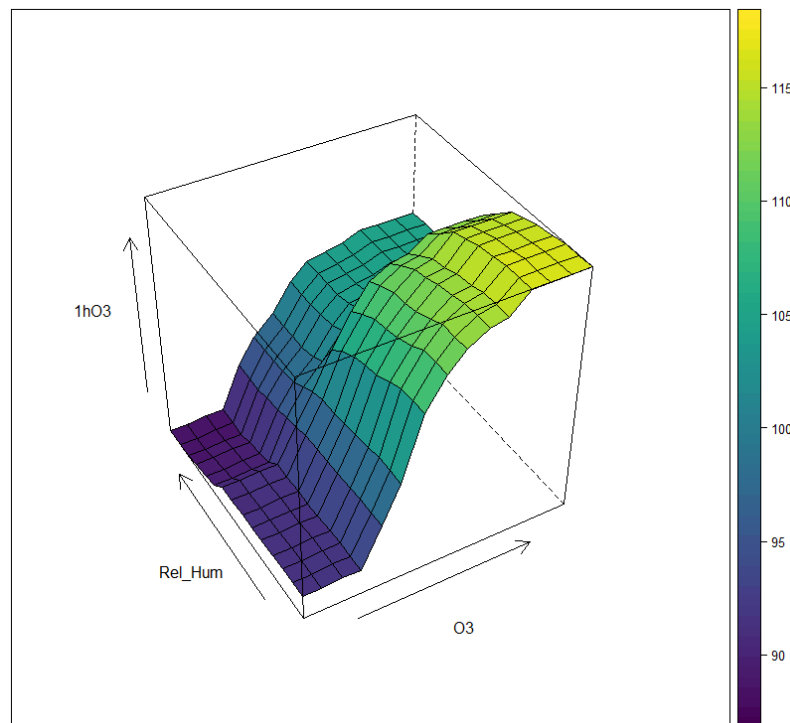


Figure 55 1hO₃ variation with respect to Rel Hum and O₃ for PR_ZU using data from May to September

5.2.2 Categorical model

Prequential evaluation results for the most accurate parameters in both couples of stations for the categorical models in days from May to September are shown in Table 25. The highest error rate is obtained for 8hO₃ in EI_RA, the lowest corresponds to 1hO₃ in PR_ZU. We had to modify the window in the prequential evaluation for 8hO₃ in PR_ZU because we did not have values for the first class with the window of the original methodology (section 4.3), the new window taken was 520 days for training, 120 for testing and 60 days for the growing block.

Output	Station	<i>n</i> tree	<i>m</i> try	Error rate
1hO ₃	PR_ZU	300	10	0.1796
8hO ₃	PR_ZU	200	12	0.2472
1hO ₃	EI_RA	300	12	0.2315
8hO ₃	EI_RA	300	9	0.4981

Table 25 Summary of the results of the prequential evaluation for the most accurate combination of parameters for every categorical model of days from May to September

The higher error rates are related to the categories with few values as we saw previously in section 5.1.2. When we took only summer days, PR_ZU concentrates a vast majority of its 1hO₃ observations over or equal to 86 µg/m³ (second category) generating high error rates for the first category in either OOB or the testing set (0.73 and 0.92 respectively), contrary to what occurs in EI_RA with error rates of 0.03 and 0.07 for OOB and testing set respectively for the first category but high error rates in the second category as we can see in Table 26. However, the error rates of the entire model of 1hO₃ are not high, the model can be very inaccurate according to the category itself.

		Confusion Matrix	1hO ₃ < 86	1hO ₃ ≥ 86	Error rate				
			Error rate of class 1	Error rate of class 2					
Barcelona-Palau Reial and Barcelona-Zona Universitaria	OOB	<table border="1" style="margin: auto;"> <tr><td>37</td><td>102</td></tr> <tr><td>31</td><td>603</td></tr> </table>	37	102	31	603	0.7338	0.0489	0.1721
	37	102							
31	603								
Testing set	<table border="1" style="margin: auto;"> <tr><td>7</td><td>81</td></tr> <tr><td>5</td><td>238</td></tr> </table>	7	81	5	238	0.9205	0.0206	0.2598	
7	81								
5	238								
Barcelona-Eixample and Barcelona-El Raval	OOB	<table border="1" style="margin: auto;"> <tr><td>588</td><td>19</td></tr> <tr><td>142</td><td>34</td></tr> </table>	588	19	142	34	0.0313	0.8068	0.2056
	588	19							
142	34								
Testing set	<table border="1" style="margin: auto;"> <tr><td>248</td><td>19</td></tr> <tr><td>43</td><td>26</td></tr> </table>	248	19	43	26	0.0712	0.6232	0.1845	
248	19								
43	26								

Table 26 Error rate of out-of-bag samples and testing set for categorical 1hO₃ in PR_ZU and EI_RA for the model of days from May to September

In both couples of stations, the 8hO₃ observations for summer are mainly accumulated within a category, more than 86 µg/m³ for PR_ZU and between 55 and 71 µg/m³ for EI_RA (Table 27). This obviously creates low error rates for these categories in these stations but considerable high error rates for other categories. It is very difficult to have low error rates for both stations with the same categorization because of the high spatial variation of the ground-level ozone. In Table 27, the general error rates of the models for both couples of stations show higher errors than the models of 1hO₃ reaching more than 0.48 similar to what we obtained in the prequential evaluation.

		Confusion Matrix	0≤8hO ₃ <55	55≤8hO ₃ <71	71≤8hO ₃ <86	8hO ₃ ≥86	Error rate
			Error rate of class 1	Error rate of class 2	Error rate of class 3	Error rate of class 4	
Barcelona-Palau Reial and Barcelona-Zona Universitaria	OOB	$\begin{vmatrix} 0 & 0 & 4 & 0 \\ 0 & 0 & 46 & 16 \\ 0 & 2 & 85 & 125 \\ 0 & 1 & 61 & 432 \end{vmatrix}$	1.0000	1.0000	0.5991	0.1255	0.3303
	Testing set	$\begin{vmatrix} 0 & 0 & 5 & 3 \\ 0 & 0 & 30 & 11 \\ 0 & 0 & 51 & 71 \\ 0 & 0 & 17 & 143 \end{vmatrix}$	1.0000	1.0000	0.5820	0.1063	0.4139
Barcelona-Eixample and Barcelona-El Raval	OBB	$\begin{vmatrix} 79 & 118 & 1 & 0 \\ 45 & 299 & 22 & 0 \\ 5 & 140 & 23 & 2 \\ 0 & 29 & 20 & 0 \end{vmatrix}$	0.6010	0.1831	0.8647	1.0000	0.4879
	Testing set	$\begin{vmatrix} 17 & 62 & 0 & 0 \\ 11 & 144 & 11 & 0 \\ 0 & 52 & 11 & 0 \\ 0 & 14 & 11 & 2 \end{vmatrix}$	0.7848	0.1325	0.8254	0.9259	0.4806

Table 27 Error rate of out-of-bag samples and testing set for categorical 8hO₃ in PR_ZU and EI_RA for the model of days from May to September

cforest delivers results about the importance analysis, which deserve to be analysed carefully because of the lower importance attributed to the solar radiation in the 1hO₃ model for PR_ZU (Figure 56). This contrasts notoriously with categorical models of the whole year and regression models of summer for the same output in the same couple of stations as well as with 8hO₃ for the same model, where Solar Rad is one of the main variables. However, Other variables such as day of the year, month, Av DPD, and Rel Hum preserve their importance with respect to the regression models (Figure 50). About pollutants, NO₂ and its moving averages are generally the most important ones for model of 1hO₃ and 8hO₃ (Figure 56) in PR_ZU. Although, there is no considerable tendency to high importance of the moving averages respect to the original daily observation.

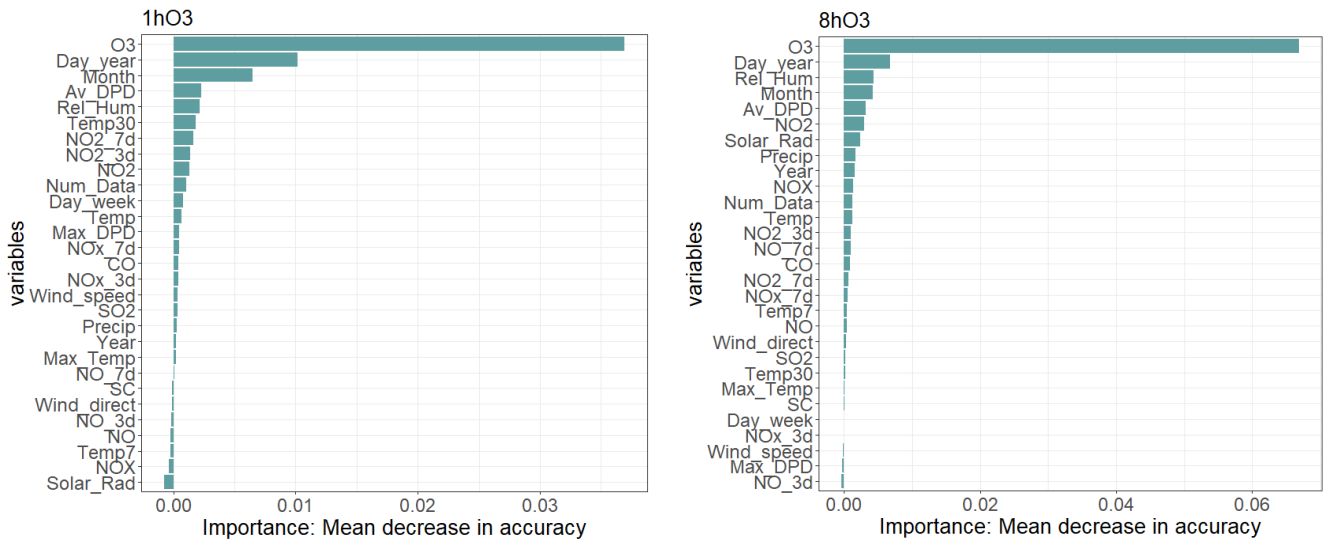


Figure 56 Variable importance for 1hO₃ and 8hO₃ in RF categorical models of days from May to September in PR_ZU

Similar to the variable importance results that we obtained in the regression models for days from May to September in EI_RA, weekday is again the second main variable after O₃ (Figure 57). However, solar radiation is not so relevant for 1hO₃ categorical model in EI_RA. On the other hand, there are several variables, which respective importance vary considerably from 1hO₃ to 8hO₃ model in EI_RA, the case of the year, solar radiation, number of data, Av DPD. These variables have a notorious change in their importance from 1hO₃ to 8hO₃. We have not seen so clearly this situation in the previous models. In every model for both couples of stations, SC has a low variable importance.

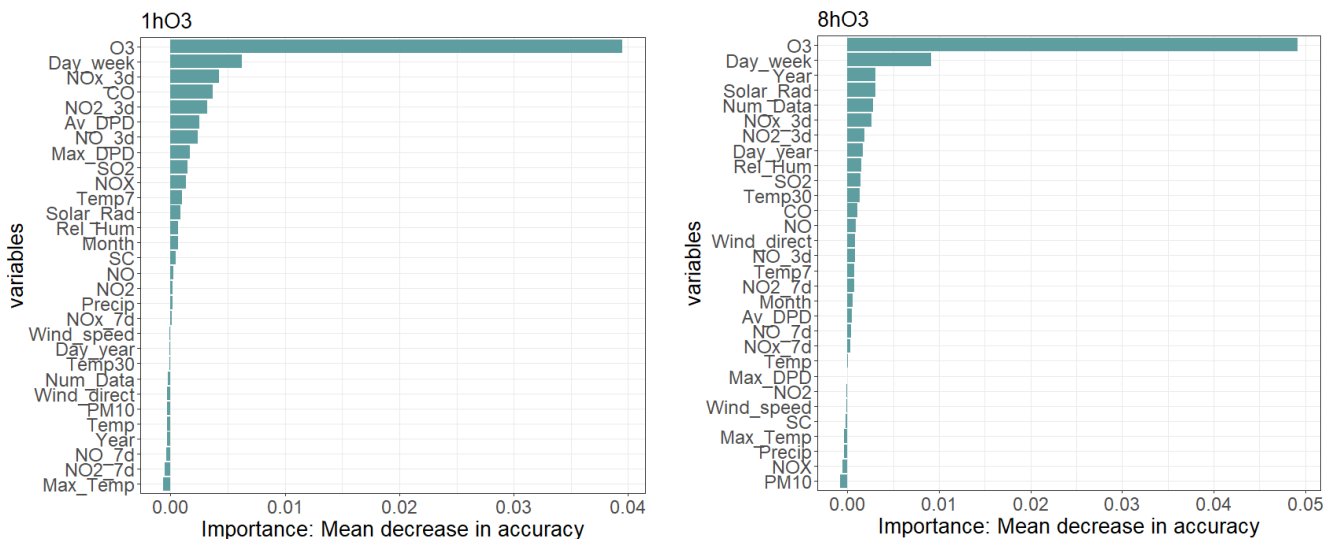


Figure 57 Variable importance for 1hO₃ and 8hO₃ in RF categorical models of days from May to September in EI_RA

We run the model without O₃ in PR_ZU once again to study the possible changes. The total error rate has a meaningless variation with respect to the complete model as we appreciate in Table 26 and Table 28. From 0.1721 to 0.1746 for OOB samples and from 0.2598 to 0.2659 for

testing set. The error rate by categories is also close to each other for the model with all variables and the model without O₃.

Barcelona-Palau Reial and Barcelona-Zona Universitaria	OOB	Confusion Matrix	1hO ₃ < 86	1hO ₃ ≥ 86	Error rate			
			Error rate of class 1	Error rate of class 2				
		<table border="1" style="margin-left: auto; margin-right: auto;"> <tr><td>17</td><td>122</td></tr> <tr><td>13</td><td>621</td></tr> </table>	17	122	13	621	0.8777	0.0205
17	122							
13	621							
Testing set	<table border="1" style="margin-left: auto; margin-right: auto;"> <tr><td>4</td><td>84</td></tr> <tr><td>4</td><td>239</td></tr> </table>	4	84	4	239	0.9545	0.0165	0.2659
4	84							
4	239							

Table 28 Error rate of out-of-bag samples and testing set for categorical 1hO₃ in PR_ZU for the model of days from May to September without O₃

Without O₃, the categorical model of 8hO₃ from May to September in PR_ZU presents a small increase in general error rate (Table 29) with respect to the model with all the variables (Table 27). However, the error rate for values of the third category (from 71 to 86 µg/m³) in the model without O₃ increases considerably.

Barcelona-Palau Reial and Barcelona- Zona Universitaria	OOB	Confusion Matrix	0 ≤ 8hO ₃ < 55	55 ≤ 8hO ₃ < 71	71 ≤ 8hO ₃ < 86	8hO ₃ ≥ 86	Error rate															
			Error rate of class 1	Error rate of class 2	Error rate of class 3	Error rate of class 4																
		<table border="1" style="margin-left: auto; margin-right: auto;"> <tr><td>0</td><td>0</td><td>3</td><td>1</td></tr> <tr><td>0</td><td>0</td><td>27</td><td>35</td></tr> <tr><td>0</td><td>0</td><td>53</td><td>159</td></tr> <tr><td>0</td><td>0</td><td>41</td><td>453</td></tr> </table>	0	0	3	1	0	0	27	35	0	0	53	159	0	0	41	453	1.0000	1.0000	0.7500	0.0830
0	0	3	1																			
0	0	27	35																			
0	0	53	159																			
0	0	41	453																			
Testing set	<table border="1" style="margin-left: auto; margin-right: auto;"> <tr><td>0</td><td>0</td><td>4</td><td>4</td></tr> <tr><td>0</td><td>0</td><td>6</td><td>35</td></tr> <tr><td>0</td><td>0</td><td>16</td><td>106</td></tr> <tr><td>0</td><td>0</td><td>5</td><td>155</td></tr> </table>	0	0	4	4	0	0	6	35	0	0	16	106	0	0	5	155	1.0000	1.0000	0.8689	0.0313	0.4834
0	0	4	4																			
0	0	6	35																			
0	0	16	106																			
0	0	5	155																			

Table 29 Error rate of out-of-bag samples and testing set for categorical 8hO₃ in PR_ZU for the model of days from May to September without O₃

Solar radiation is one of the main variables again for the categorical model of 1hO₃ without O₃ for days from May to September for PR_ZU (Figure 58). In both models, either 1hO₃ or 8hO₃, we have similar variables as the main ones, i.e., day of the year, relative humidity, NO_{2_7}, solar radiation and similar variables as the less important ones. In general terms, the order of variable importance has not significantly changed when we remove O₃, except for solar radiation.

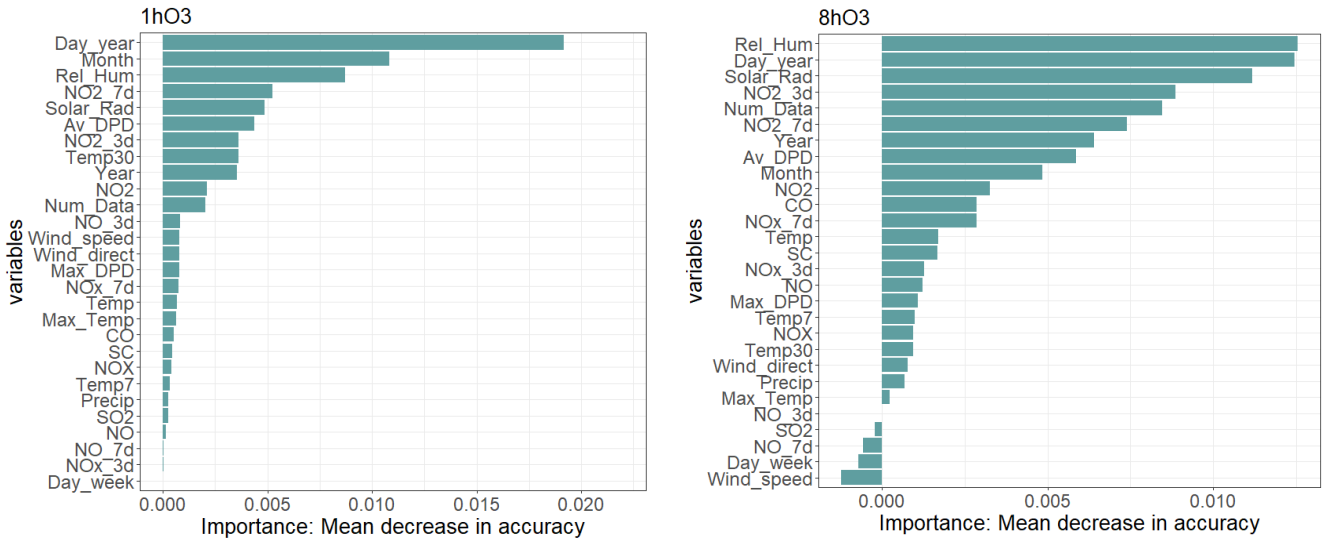


Figure 58 Variable importance for 1hO₃ and 8hO₃ in RF categorical model of days from May to September without O₃ in PR_ZU

6 Conclusions

6.1 Regression models

The RF regression model with data of the whole year is capable of capturing the time variation of the daily maximum hourly ozone concentration level ($1hO_3$) and the daily maximum 8-hours average ozone concentration level ($8hO_3$) for both couples of stations in Barcelona as we could see in the time series plots and observed vs predicted scattered plots where the point cloud of the testing set follows approximately the ideal tendency. This situation does not occur when we consider data only from May to September where the models are not able to capture consistently the time variation of the testing set for both outputs.

Taking into account the magnitude of the values ground-level ozone where the range of $1hO_3$ for the whole year corresponds to 222 and 208 [$\mu\text{g}/\text{m}^3$] for PR_ZU and EI_RA respectively, and the ranges of $8hO_3$ are 164.75 and 141.75 [$\mu\text{g}/\text{m}^3$] in the same order, the error metrics such as RMSE (from 11.2 to 13.4 [$\mu\text{g}/\text{m}^3$]) or MAE (from 8.78 to 9.97 [$\mu\text{g}/\text{m}^3$]) for both outputs show us that the model has an acceptable approximation to the observations. This can be also appreciated in the results of the mean absolute percentage error (MAPE) for the whole year where only the EI_RA model of $8hO_3$ presents a value over 23% and the rest of the models are below 20%. Whole year models have an admissible accuracy. Similar error metrics are reached for the models from May to September achieving even lower MAPE values (<15%). However, summer models cannot capture temporal variation consistently. Hence, these models have low accuracy. The error metrics are considered from the testing set and OOB analysis in every case, where both are always similar for the regression models.

The addition of Jenkinson and Collison synoptic classification as a categorical variable into the regression models does not improve the accuracy of the models notoriously. Moreover, this variable is not part of the main variables in the models. However, pure advection and anticyclonic advection coming from the east are the classifications related to high ozone concentration values. The characterization of the atmospheric circulation based on another classification even considering a smaller scale (not taking all Catalonia under a unique classification) might be useful.

O_3 ($1hO_3$ or $8hO_3$ of the previous day to the prediction) is the main variable in every RF regression model for either the whole year or only with days from May to September. This has a positive correlation with the predicted values. O_3 values usually vary by a small magnitude with respect to the next day (O_3 values have some inertia). Therefore, O_3 can be taken as an approximation for the values of the next day. However, there are differences in the rest of the main variables between the models of the whole year and only summer days (May to September).

In the models of the whole year, solar radiation, day of the year, the moving average of 7 days of NO and months are the main variables after O_3 for both stations. The model captures correctly the behaviour of the time variables. They have a positive correlation with the predicted values from January to summer months and from there to December, the correlation is negative. Solar

radiation has a positive correlation, which goes according to the reality considering how ozone is generated. NO and its daily moving averages show a negative correlation with the predicted values, and at the same time, NO shows a decrease during weekends where ozone concentration levels increase. Consequently, the negative variation of this variable might explain the high values of ground-level ozone during weekends. In addition, NO acquires higher importance while the period of the moving average is wider.

Days of the week is one of the main variables for EI_RA in the whole year and with days from May to September indicating that the traffic reduction has a higher influence in the city centre than in the surrounding areas. This characteristic of the model represents correctly the reality in the study area.

In the models with data from May to September, relative humidity and average dew-point deficit are part of the main variables for both outputs (more important for 1hO₃ than 8hO₃) in contrast to the models of the whole year. These two variables have a higher importance in the areas far from the city centre. NO₂ increases its importance with respect to the models of the whole year, this can be appreciated in areas close and far from the city centre. On the other hand, NO decreases its importance. The moving average of three days of NO₂ and NO_x acquire high importance for both outputs in the city centre consolidating nitrogen oxides (NO, NO₂ and NO_x) as the most important air quality variables in the models, this represents correctly the nature of the ground-level ozone. In the case of NO₂, there is a positive correlation with respect to ozone concentration levels. There is no clear importance superiority of the moving averages over daily measurements in these models. The main differences in the variable importance analysis between the models of the whole year and summer are related to the nature of the models. While summer models take mainly high ground-level ozone values, models of the whole year consider the full range of variation.

randomForest package produces a considerable bias increasing the importance of the categorical variables, especially when the importance is measured based on MDI. *cforest* package delivers more reliable results about importance, and with a similar accuracy. Hence, it is recommended to use this package when we incorporate categorical variables such as SC to the model.

6.2 Categorical models

There are very few events that overcome the information and alert limits established by the Directive 2008/50/CE in both 1hO₃ and 8hO₃. Consequently, it is not possible to train a model based on those categories. In order to categorize the outputs of the model, we selected a category based on the standards given by EPA. These thresholds allow us to have several values in every category. However, the distribution of the values is not uniform for every category. Therefore, some categories gather a large number of values, and others include too few, creating high error rates in some categories. The same categorization for every couple of stations gives a different distribution of the values. Hence, there is no optimal and unique classification to obtain a uniform distribution of the outputs.

In general terms, categorical models for 1hO₃ show low error rates and for 8hO₃, error rates are higher. However, when we individually see the error rates for every category inside the models, we find that the categories with a lot of values inside them have considerable low errors but categories with few values have high error rates, and we have the same scenario in every model and in both couples of stations. Therefore, these models are not useful to alert the population about a specific category, which might be dangerous for human health. Further analysis with modified datasets, which allow us to have the same number of values in every category might be considered.

As we had in the regression models, O₃ is the main variable for every categorical model. The same variables that are the most important ones for the regression models of the whole year are the main variables for the categorical models for the same period of time in both couples of stations. Temperature variables acquire high importance in PR_ZU but not in EI_RA. On the other hand, weekday is the second main variable in EI_RA indicating again the high importance of the traffic variation in the city centre.

cforest shows some difficulties determining the importance of the variables in the categorical models of days from May to September because one of the most important variables in the rest of the models (solar radiation) is catalogued as the least important for the 1hO₃ model of PR_ZU. This is not related to the chemistry of the generation of ozone, and even when we remove O₃ from the dataset, this variable changes its importance considerably, which is not consistent. However, the most important variables in the regression models for days from May to September have also the main positions in the categorical models with the same differences between locations of the stations that we saw in the previous section.

References

Abdul-Wahab, S.A., Al-Alawi, S.M., (2002) 'Assessment and prediction of tropospheric ozone concentration levels using artificial neural network', *Environmental Modelling & Software*, 17, pp. 219-228, [https://doi.org/10.1016/s1364-8152\(01\)00077-9](https://doi.org/10.1016/s1364-8152(01)00077-9)

Agirre, E., Anta, A., Barrón, L.J. R., Albizu, M., (2007) 'A neural network based model to forecast hourly ozone levels in rural areas in the Basque Country', *WIT Transactions on Ecology and the Environment*, Air Pollution XV, 101, pp. 109-118, <https://doi:10.2495/AIR070111>

Akimoto, H., Izuta, T., Ueda, H., Uchiyama, I., Ohara, T., Kohno, Y., Kobayashi, K., Wakamatsu, S., (2006) *Tropospheric Ozone A Growing Threat*, Acid Deposition and Oxidant Research Center.

Ajuntament de Barcelona, (2020) Estadística i Difusió de Dades, Tipologia del parc de vehicles, available at:
https://ajuntament.barcelona.cat/estadistica/catala/Estadistiques_per_temes/Transport_i_mobilitat/Mobilitat/Vehicles/Parc_de_vehicles/a2020/tipo/t01.htm

Aljanabi, M., Shkoukani, M., Hijjawi, M., (2020) 'Ground – level Ozone Prediction Using Machine Learning Techniques: A Case Study in Amman, Jordan', *International Journal of Automation and Computing*, 17 (5), pp. 667-677, <https://doi:10.1007/s11633-020-1233-4>

Alves, L., Nascimento, E., Moreira, D., (2019) 'Hourly Tropospheric Ozone Concentration Forecasting Using Deep Learning', *WIT Transactions on Ecology and Environment*, 236, pp. 129-138, <https://doi:10.2495/AIR190131>

Breiman, L., (2001) 'Random Forest', *Machine Learning*, 45(1), 5-32, <https://doi.org/10.1023/A:1010933404324>

Cerqueira, V., Torgo, L., Mozetič, I. (2020) 'Evaluating time series forecasting models: an empirical study on performance estimation methods', *Machine Learning*, 109(11), pp.1997–2028. <https://doi.org/10.1007/s10994-020-05910-7>

Chameides, W.L., Fehsenfeld, F., Rodgers, M.O., Cardelino, C., Martinez, J., Parrish, D., Lonneman, W., Lawson, D. R., Rasmussen, R. A., Zimmerman, P., Greenberg, J., Middleton, P., Wang, T., (1992) 'Ozone Precursor Relationship in the Ambient Atmosphere', *Journal of Geophysical Research*, 97, pp 6037-6055, <https://doi.org/10.1029/91jd03014>

Chong, S.F., & Choo, R., (2011) 'Introduction to Bootstrap', *Proceeding of Singapore Healthcare*, 20, pp. 236-240, <https://doi.org/10.1177/201010581102000314>

European Commission, (2013) *Decision 2011/850/EU*, European Parliament and the Council of the European Union, Available at:

https://ec.europa.eu/environment/air/quality/legislation/pdf/IPR_guidance1.pdf

Feng, R., Zheng, H., Zhang, A., Huang, C., Gao, H., Ma, Y., (2019) 'Unveiling tropospheric ozone by the traditional atmospheric model and machine learning, and their comparison: A case study in Hangzhou, China', *Environmental Pollution*, 252, pp. 366 – 378, <https://doi.org/10.1016/j.envpol.2019.05.101>

Ghojogh, B., Crowley, M., (2019) *The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Booting: Tutorial*.

Haq, G., Schwela, D., (2008) *Foundation Course on Air Quality Management in Asia – Monitoring*, Stockholm Environment Institute.

Hothorn, T., Hornik, K., Strobl, C., Zeileis, A., (2021) *cluster: party*, R package version 1.3-7, A Laboratory for Recursive Partytioning, <http://party.r-forge.r-project.org/>

IDESCAT, (2020) Available at: <https://www.idescat.cat/emex/?id=080193&lang=es>.

INE, (2016) Instituto Nacional de Estadística – Boletín Mensual de Estadística. Diciembre, available at: <https://www.ine.es/daco/daco42/bme/c19.pdf>

Inicio de la campaña de vigilancia de ozono troposférico, (2021) Departamento de territorio y sostenibilidad, Generalitat de Catalunya, Available at: <http://mediambient.gencat.cat/es/details/Noticies/20210515-inici-campanya-ozo>

Ioannou, K., Karampatzakis, D., Amanatidis, P., Aggelopoulos, V., Karmiris, I., (2021) 'Low-Cost Automatic Weather Stations in the Internet of Things', *Information*, MDPI, 12, 146, <https://doi.org/10.3390/info12040146>

Krzyzanowski, M., Cohen, A., (2008) 'Update of WHO air quality guidelines', *Air Quality Atmospheric Health*, (1)7 – 13, <https://doi.org/10.1007/s11869-008-0008-9>

Lantz, B., (2019) *Machine Learning with R, Expert techniques for predictive modeling*, Packt Publishing, Third Edition.

Liaw, A., & Wiener, M., (2018) *cluster: randomForest*, R package version 4.6-14, Breiman and Cutler's Random Forest for Classification and Regression, Based on Leo Breiman and Adele Culter original Fortran, <https://www.stat.berkeley.edu/~breiman/RandomForests/>

Luna, A.S., Paredes, M.L.L., de Oliveira, G.C.G., Correa, S.M., (2014) 'Prediction of ozone concentration in tropospheric levels using artificial neural networks and support vector machine at Rio de Janeiro, Brazil', *Atmospheric Environment*, 98, pp. 98-104, <http://dx.doi.org/10.1016/j.atmosenv.2014.08.060>

Malinović-Milićević, S., Vykylyuk, Y., Stanojević, G., Radovanović, M., Doljak, D., Čurčić, N (2021) 'Prediction of tropospheric ozone concentration using artificial neural network at traffic and background urban locations in Novi Sad, Serbia', *Environment Monitoring and Assessment*, 193:84, <https://doi.org/10.1007/s10661-020-08821-1>

Martín-Vide, J., Moreno García, M. C., Artola, V. M., & Cordobilla, M. J. (2016) 'Los tipos sinópticos de Jenkinson & Collison y la intensidad de la isla de calor barcelonesa'. *Clima, sociedad, riesgos y ordenación del territorio*, pp. 565–573. <https://doi.org/10.14198/xcongresoaealicante2016-53>

McKeen, S. A., Hsie, E. Y., & Liu, S. C. (1991) 'A study of the dependence of rural ozone on ozone precursors in the eastern United States', *Journal of Geophysical Research*, 96(D8), pp. 15,377-15,394, <https://doi.org/10.1029/91jd01282>

Meng, Z.Y. (2019) 'Ground Ozone Level Prediction Using Machine Learning', *Journal of Software Engineering and Applications*, 12, pp. 423-431. <https://doi.org/10.4236/jsea.2019.1210026>

Naser, M.Z. Alavi, A. H., (2020) *Insights into Performance Fitness and Error Metrics for Machine Learning*.

Nwanganga, F., Chapple, M., (2020) *Practical Machine Learning in R*, Wiley.

Oliveira, M., Torgo, L., Santos Costa, V. (2021) 'Evaluation Procedures for Forecasting with Spatiotemporal Data', *Mathematics*, 9(6), 691, <https://doi.org/10.3390/math9060691>

Pernak, R., Alvarado, M., Lonsdale, C., Mountain, M., Hegarty, J., Nehr Korn, T., (2019) 'Forecasting Surface O₃ in Texas Areas Using Random Forest and Generalized Additive Models', *Aerosol and Air Quality Research*, 19, pp. 2815-2826, <https://doi.org/10.4209/aaqr.2018.12.0464>

Potdar, K., Pardawala, T., Pai, C.D., (2017) 'A Comparative Study of Cartegorical Variable Encoding Techniques for Neural Network Classifiers', *International Journal of Computer Applications*, 175 (4), <https://doi.org/10.5120/ijca2017915495>

Sensirion, (2006) Application Note Dew-point calculation, SHTxx Humidity and Temperature, available at:
http://irtfweb.ifa.hawaii.edu/~tcs3/tcs3/Misc/Dewpoint_Calculation_Humidity_Sensor_E.pdf

Sillman, S., (1993) 'Tropospheric Ozone: The Debate over Control Strategies', *Annual Reviews of Energy Environment*, Vol 18, pp.31-56, <https://doi.org/10.1146/annurev.eg.18.110193.000335>

SINDIC, (2019) *La calidad de aire en Cataluña: Déficits y Recomendaciones*, Available at: http://www.sindic.cat/site/unitFiles/6328/Informe%20qualitat%20aire_cast_def.pdf

Soriano, C., Baldasano, J.M., Buttler, W., Moore, K.R., (2001) 'Circulatory Patterns of Air Pollutants within the Barcelona Air Basin in a Summertime Situation: Lidar and Numerical Approaches', *Boundary-Layer Meteorology*, 98, pp. 33-55. <https://doi.org/10.1023/a:1018726923826>

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007) 'Bias in random forest variable importance measures: Illustrations, sources and a solution', *BMC Bioinformatics*, 8(1). <https://doi.org/10.1186/1471-2105-8-25>

Texas Commission on Environmental Quality, (2018) *High Ozone in your Metro Area*, Available at: https://www.tceq.texas.gov/cgi-bin/compliance/monops/ozone_summary.pl#interpret

The European Parliament and the Council of the European Union, (2008) *Directive 2008/50/CE of European Union*.

UCAR Centre for Science Education, (2021) *Diagram of Atmosphere Layers*, <https://scied.ucar.edu/image/atmosphere-layers-diagram>

Wang, N., Lyu, X., Deng, X., Huang, X., Jiang, F., & Ding, A. (2019) 'Aggravating O₃ pollution due to NO_x emission control in eastern China', *Science of The Total Environment*, 677, pp. 732–744, <https://doi.org/10.1016/j.scitotenv.2019.04.388>

Wilson, R. C., Fleming, Z. L., Monks, P. S., Clain, G., Henne, S., Konovalov, I. B., Szopa, S., & Menut, L. (2012) 'Have primary emission reduction measures reduced ozone across Europe? An analysis of European rural background ozone trends 1996–2005', *Atmospheric Chemistry and Physics*, 12(1), 437–454. <https://doi.org/10.5194/acp-12-437-2012>

World Bank Group, (1998) 'Ground-Level Ozone', *Pollution Prevention and Abatement Handbook*.

Xu, J., Zhang, Y., Miao, D. (2020). 'Three-way confusion matrix for classification: A measure driven view', *Information Sciences*, 507, pp. 772–794. <https://doi.org/10.1016/j.ins.2019.06.064>

Ying, X., (2019) 'An Overview of Overfitting and its Solutions', *Journal of Physics: Conference Series*, 1168, <https://doi.org/10.1088/1742-6596/1168/2/022022>

Zabkar, R., Zabkar, J., Cemas, D., (2004) 'Ground-level ozone forecast based on machine learning', *Air Pollution XII*, WIT press, pp. 41-48.