

Article

Forecasting Daily of Surface Ozone Concentration in the Grand Casablanca Region Using Parametric and Nonparametric Statistical Models

Halima Oufdou ^{1,*} , Lise Bellanger ², Amal Bergam ³ and Kenza Khomsi ⁴

¹ Laboratory of Applied Economics, Agdal FSJES, Mohammed V University, B.P. 721, Rabat 10056, Morocco

² Laboratory of Mathematics Jean Leray UMR CNRS 6629, University of Nantes, 44322 Nantes, France; lise.bellanger@univ-nantes.fr

³ Laboratory MAE2D, University of Abdelmalek Essaadi, Larache 92004, Morocco; abergam@uae.ac.ma

⁴ National Climate Center, Air Quality Department, General Directorate of Meteorology, B.P. 8106, Casablanca 20000, Morocco; khomsi.kenza@marocmeteo.ma

* Correspondence: h.oufdou@um5r.ac.ma

Abstract: Forecasting concentration levels is important for planning atmospheric protection strategies. In this paper, we focus on the daily average surface ozone (O_3) concentration with a short-time resolution (one day ahead) in the Grand Casablanca Region of Morocco. The database includes previous day O_3 concentrations measured at Jahid station and various meteorological explanatory variables for 3 years (2013 to 2015). Taking into account the multicollinearity problem in the data, adapted statistical models based on parametric (SPLS and Lasso) and nonparametric (CART, Bagging, and RF) models were built and compared using the coefficient of determination and the root mean square error. We conclude that the parametric models predict better than nonparametric ones. Finally, from the explanatory variables stored by the SPLS and Lasso parametric models, we deduce that a very simple linear regression with five variables remains the most appropriate for the available data at Jahid station ($R^2 = 0.86$ and $RMSE = 9.60$). This resulting model, with few explanatory variables to prevent missing data, has good predictive quality and is easily implementable. It is the first to be built to predict ozone pollution in the Grand Casablanca region of Morocco.

Keywords: air pollution; tropospheric ozone; meteorological variables; Morocco; forecast; statistical models



Citation: Oufdou, H.; Bellanger, L.; Bergam, A.; Khomsi, K. Forecasting Daily of Surface Ozone Concentration in the Grand Casablanca Region Using Parametric and Nonparametric Statistical Models. *Atmosphere* **2021**, *12*, 666. <https://doi.org/10.3390/atmos12060666>

Academic Editors: Rafaella Eleni P. Sotiropoulou and Efthimios Tagaris

Received: 30 April 2021

Accepted: 20 May 2021

Published: 23 May 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Over the past decades, several studies have been developed, clearly showing the impact of air pollution on human health [1,2], the environment, the natural resources, and the sustainable development of many regions [3]. Morocco is one of the countries with an arid or semi-arid climate, especially those on the southern shore of the Mediterranean, are exposed to air pollution. The barren soils and the high temperature give rise to high O_3 emissions due not only to automobile traffic and industrialization but also to significant soil contributions, linked to the aridity of the climate and the proximity of the desert [4]. Therefore, Morocco is not far from this deterioration of ecological conditions, particularly in large areas where 13.4% of Moroccan population and most the human activities are accumulated (Industries, vehicles, etc.) such as the Grand Casablanca Region (GCR) [5]. O_3 is a secondary trace gas in the atmosphere, and it is not directly emitted from a natural or anthropogenic source but rather formed by a complex set of chemical reactions involving nitrogen oxides (NO_x), carbon monoxide (CO), methane (CH_4), and volatile organic compounds (VOC) in the presence of the sun [3]. Surface O_3 concentrations are also influenced by meteorological conditions that have a significant role in the transport of concentrations of this pollutant such as temperature, pressure, humidity, wind direction, sunshine duration, etc. [4,6,7]. O_3 threatens human health and environment [3,8]. In fact,

epidemiological studies have shown that current ambient exposures are associated with reduced basic pulmonary function, exacerbation of asthma, and premature mortality [1,9]. To avoid this problem, the prediction of O₃ concentrations remains a crucial and necessary step in controlling pollution and in mitigating its adverse effects. However, the series of chemical reactions before the emission of ozone into the troposphere during the formation process complicates its forecast [10,11].

In the absence of a statistical forecast model to predict daily ozone from one day to the next for the most critical air pollutants in GCR, the current study is considered the first of its kind in Morocco. The statistical approach is based on historical data in order to predict the future behavior of O₃ associated in large part with meteorological conditions [12]. During the last decade, many researchers have studied the problem of forecasting tropospheric ozone using multivariate statistical methods. These methods can be classified into two main categories: parametric and nonparametric methods [13–15]. Recent studies, such as [16–21], have compared the performance of these forecasting methods in selecting the most appropriate one. The results have shown that Multiple Linear Regression (MLR), regression tree, and Random Forests give good results. However, these methods are still associated with disadvantages that make interpretation more difficult such as the following: (i) the MLR retains a large number of predictors, which often present multicollinearity problems solved in our last paper by comparing nine alternative regression models [22], and (ii) the construction process of the regression tree and random forests methods remains complex [23]. In addition, these methods require more time during their development and are not recommended when the data history is limited with missing data (case of GCR).

In this regard, the paper at hand proposes a new model that can be easily implemented and provides predictions of O₃ concentrations with a reduced number of explanatory variables that retain good predictive qualities. This model is based on a reduced data history of 3 years (2013, 2014, and 2015) containing missing data. The final model was built and validated by the following process: (i) development of the model in the data from 2013 and 2014 by introducing observed meteorological variables and (ii) validation of the model using data from 2015 by injecting predicted meteorological variables. The results obtained are very satisfactory in terms of the predictive capacity compared to the other models in the present study.

The article at hand is organized as follows: Section 2 is devoted to the study's area, the description and analysis of the data, the brief description of the parametric and nonparametric statistical models compared, and the model evaluation. Section 3 sheds light on the results obtained, and Section 4 provides a general discussion with perspectives.

2. Materials and Methods

2.1. Study Area and Data Collection

Morocco is an African country located in the extreme northwest of the continent. It is located in the southern part of the Mediterranean basin and is considered among the most vulnerable countries to climate variability and trans-boundary air pollution [24]. The Grand Casablanca (Figure 1) is the studied region (33°34'42.44" North, 7°36'23.89" West), is located in the central-western part of the Kingdom of Morocco on the Atlantic Ocean coast, and covers an area of 1117 km². The Grand Casablanca Region (GCR) has experienced rapid expansion and increased population growth. It concentrates nearly 13.4% of the total population of the country. Its population is estimated at 4,270,750 residents in 2014 according to World Population Review web-page [5]. In addition, GCR is also one of the most important regions in Africa because it is seen as a hub of economic and business activity. The climate is Mediterranean with a strong oceanic tendency. Temperature can rise up to 40 °C. During the day, the region suffers from mixed episodes (alternating ocean winds, offshore breezes, and synoptic flow) resulting from anticyclonic type conditions that translate into a northeasterly flow [5]. In general, air pollution by the O₃ pollutant depends on several meteorological parameters and the modeling of this phenomenon requires a complete study of the available database. The data over the study area were provided

by the General Directorate of Meteorology (GDM) as a result of a scientific cooperation agreement between the GDM and the Poly-disciplinary Faculty of Larache, where this work was conducted. The historical records of these data cover a 3-year period: from 1 January 2013 to 31 December 2015 corresponding to (i) observed meteorological data measured by the meteorological station; (ii) next day's forecast meteorological data corresponding to the outputs from the numerical meteorological forecast model called "Albahir", which is based on the community model ALADIN (Limited Area Dynamic Adaptation International Development) [24]; and (iii) O₃ concentrations measured at the Jahid station located in the center of Casablanca City (Figure 1) and having the most important historical record compared to other stations within the region. All statistical analyses were performed using the free software R (<http://www.r-project.org> (accessed on 1 April 2021)).

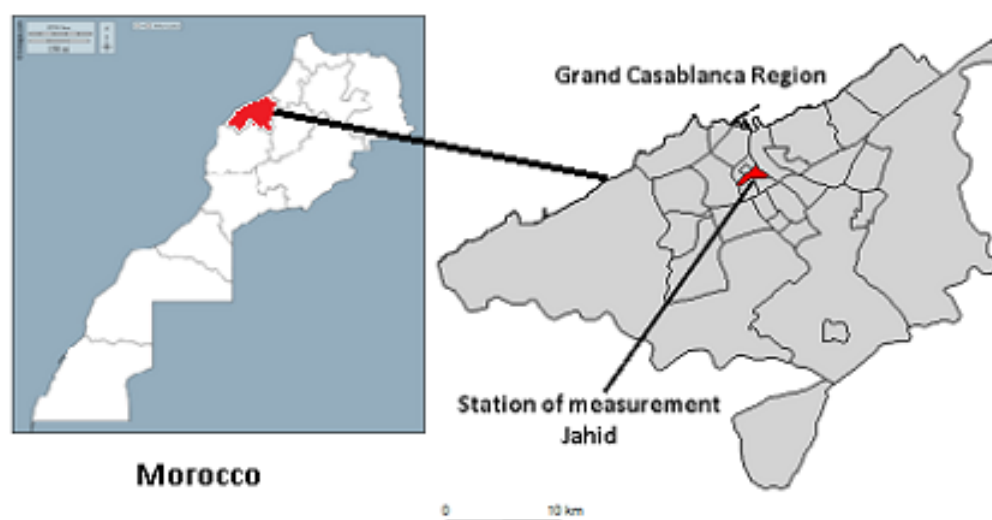


Figure 1. Map of GCR. The measurement station is a urban station (Jahid) located in the center of Casablanca. Source: Global Administrative Areas (GADM).

2.2. Modelling Approach

The statistical approach adopted in this study is focused on the comparison of two types of forecasting models most frequently used in the literature: parametric (Appendix A) and nonparametric models (Appendix B). In the parametric models, two parsimonious alternative methods to the Multiple Linear Regression (MLR), Sparse Principal Least Squares (SPLS) [25], and Least Absolute Shrinkage and Selection Operator (Lasso) regression [26,27] methods were selected, which conserve a limited number of the most significant variables using penalties on the norm of the variable weights for SPLS and on the L1 norm of the coefficient vector for Lasso. A more detailed description of the two methods can be found in Appendix A.

At the level of nonparametric models, three methods often used in ozone forecasting have been studied: (i) Classification and Binary Regression Tree (CART) [28], which provides a binary regression tree facilitating the identification of the most important variables; (ii) Bootstrap Aggregating (Bagging) [29], which is a random method that allows a user to average the predictions of several independent models by reducing the variance and the prediction error under the bootstrapping principle; and (iii) Random Forests (RF) [30], which is an improvement in Bagging that allows a user to aggregate regression trees by inserting a random selection of a limited number of variables among all of the studied predictors variables. The details of these methods are given in Appendix B. Several statistical studies combining previous day's ozone and meteorological factors to predict daily O₃ concentrations have been published. Various studies in different countries have used a modeling approach based on the following models:

- parametric linear models (Multiple Linear Model and its alternative models): Brazil [31], Canada [32], China [33], Croatia [34], Greece [35], Italy [36], Portugal [14], Spain [37], Malaysia [38], the United States [39], Mexico [21], etc.
- nonparametric models (CART Tree, Bagging, and Random Forests): the United States [40], France [41], Spain [42], China [17], Sweden [43], etc.

In this context, a comparison between these different models was performed in a first step and used the best models obtained in the construction of a new model in a second step. This final regression model was qualified by a reduced number of predictor variables and was easily implementable. It is the first model for daily ozone forecasting in Morocco, more precisely in the GCR.

2.3. Model Evaluation

The evaluation of the fit and performance of the models were performed in two phases necessary to interpret and compare the results of the models, called cross-validation technique. The evaluation of the fit and performance of the models was conducted in two main steps following the cross-validation technique [12,26]. The first phase concerns the internal validation performed on the training data sets, and the second one is called external validation performed on the test data. This is a simple technique that considers the first data set as a training sample (67% of the total data) and the second sample as test or validation data (33% of the total data including observed and predicted meteorological data). The test data set is not used to develop the model but only to evaluate it. Since the present study is based on a 3 year records, the spirit of threefold cross validation has been chosen [44]. The sample is divided into three subsets of one year and for each fold. After that, the models were estimated on all subsets except one (see Figure 2). The left out subset is used to test the model and to determine the most accurate O₃ prediction model in terms of predictive capability.

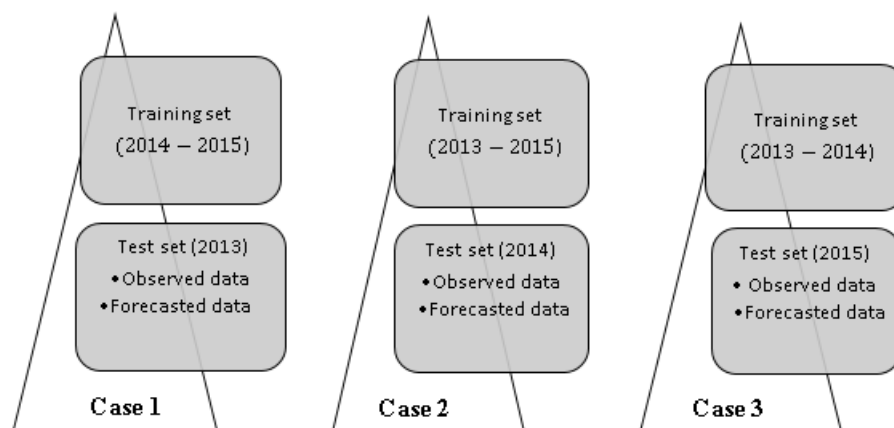


Figure 2. Scheme of the three cases tested using the spirit of k-fold cross validation technique.

After comparing the obtained results from the models tested for each case of study according to the examined three years (2013–2015), the third case according to the real chronological order was chosen (Figure 2). Indeed, this period is marked by a stability in the results of the studied models, thus ensuring a better forecast. Tables A1 and A2 give the results obtained in case 1 and case 2, respectively.

During the adjustment (internal validation) and validation (external validation) of the different models on training and test data, the comparison of their performance was evaluated with the main types of statistical criteria and calculated at each step by equations for R^2 and $RMSE$.

- The coefficient of determination denoted R^2 . This statistic (Equation (1)) provides a measure of the proportion of the variance in the response variable that is predictable

from the explanatory variables. It gives some information about the goodness of fit of a model. It ranges from 0 to 1: the closer its value is to 1 the better the model is.

$$R^2 = 1 - \frac{\sum_{i=1}^{n_{\text{train}}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{\text{train}}} (y_i - \bar{y})^2} \quad (1)$$

where n_{train} is the size of training sample, y_i is the y value for i observation, $\bar{y}$ is the mean y value, and $\hat{y}_i$ is the prediction of i observation obtained using the MLR model (Appendix A).

- The Root Mean Squared Error (RMSE) is the standard deviation of the residuals (prediction errors). This is computed according to the following expression (Equation (2)):

$$RMSE = \sqrt{\frac{1}{n_{\text{train}}} \sum_{i=1}^{n_{\text{train}}} (y_i - \hat{y}_i)^2} \quad (2)$$

The smallest value of this criterion corresponds to the best goodness of fit of the model. To assess the predictive capacity of the models, we use the RMSE criterion calculated from the observed data of summer 2015 named $RMSEP_{obs}$ (Equation (3)):

$$RMSEP_{obs} = \sqrt{\frac{1}{n_{\text{obtest}}} \sum_{i=1}^{n_{\text{obtest}}} (y_i - \hat{y}_i)^2} \quad (3)$$

where n_{obtest} is the size of the observed the validation set (obtest).

Obviously, the best predictive model corresponds to the smallest $RMSEP_{obs}$.

In the same way, the $RMSEP$ of prevision based on the real forecasted meteorological data set on 2015 named $RMSEP_{prev}$ is defined as follows (Equation (4)):

$$RMSEP_{prev} = \sqrt{\frac{1}{n_{\text{prevtest}}} \sum_{i=1}^{n_{\text{prevtest}}} (y_i - \hat{y}_i)^2} \quad (4)$$

where n_{prevtest} is the size of the sample size of the forecasted data (prevtest).

3. Results

In this section, the results obtained with two studied categories of models are presented: parametric (SPLS and Lasso) and nonparametric ones (CART, Bagging, and RF). The purpose of this study is to propose the simplest statistical model for predicting daily O_3 from day i to day $(i + 1)$ taking into account the least number of variables to prevent the problem of missing values.

3.1. Data Preparation

We dispose of 23 quantitative explanatory variables for the observed and forecasted meteorological data set including sunshine duration, temperature, relative humidity, wind speed, wind direction, pressure, etc. as well as the pollution data (ozone concentrations on the days i and $i - 1$). Table 1 below provides the variable's abbreviation and units of measurement.

Table 1. Variable abbreviations and units of measurement.

Abbreviation	Variable	Unit
TMPMAX	Maximal temperature	°C
TMPMIN	Minimal temperature	°C
TMPMOY	Average temperature	°C
RRQUOT	Total precipitation	Mm
DRINSQ	Sunshine duration	Heure
HUMREL06h	Relative humidity at 6 h	%
HUMREL12h	Relative humidity at 12 h	%
HUMREL18h	Relative humidity at 18 h	%
PRESTN06h	Pressure at the station level at 6 h	HPA
PRESTN12h	Pressure at the station level at 12 h	HPA
PRESTN18h	Pressure at the station level at 18 h	HPA
FFVM06h	Wind force at 6 h	m/s
FFVM12h	Wind force at 12 h	m/s
FFVM18h	Wind force at 18 h	m/s
DDVM06h	Wind direction at 6 h	Degree
DDVM12h	Wind direction at 12 h	Degree
DDVM18h	Wind direction at 18 h	Degree
Vx06	Horizontal wind at 6 h	m/s
Vx12	Horizontal wind at 12 h	m/s
Vx18	Horizontal wind at 18 h	m/s
Vy06	Vertical wind at 6 h	m/s
Vy12	Vertical wind at 12 h	m/s
Vy18	Vertical wind at 18 h	m/s
$O_{3\text{veilleJahid}}$	O_3 concentrations of previous day	$\mu\text{g}/\text{m}^3$
$O_{3\text{Jahid}}$	Ozone concentrations	$\mu\text{g}/\text{m}^3$

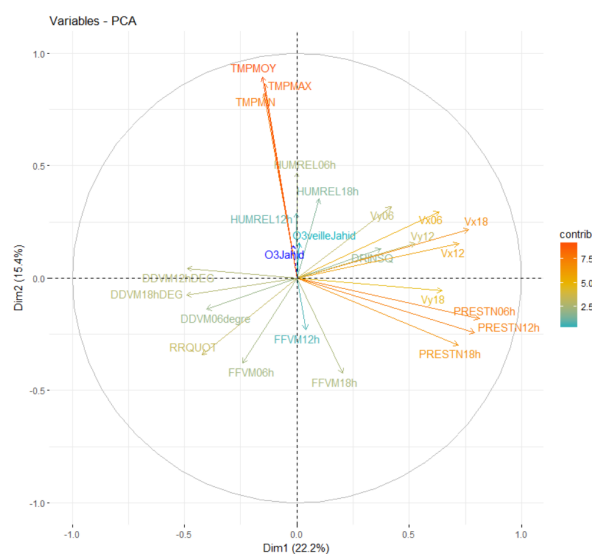
Historical data related to the three years 2013, 2014, and 2015 were measured in the GCR. Since the aim of our work is to forecast daily O_3 , which is a pollutant qualified as a secondary species and significantly behaves differently depending on the time of the year (summer or winter), the present study was limited to the summer period (April–September) for which ozone concentrations are the highest [7]. Three summer periods were considered (2013, 2014, and 2015), i.e., 549 days. Table 2 presents a description of the observed meteorological and air pollution factors according the different statistical indicators (minimum, maximum, mean, and standard deviation) and the multicollinearity factor (*VIF*). The *VIF* indicator allows us to analyze the relationship between the predictor variables by using MLR model. The multicollinearity problem is identified as soon as $VIF < 5$ [45]).

Indeed, the concentration of O_3 during these years is between 10 and 130 $\mu\text{g}/\text{m}^3$, the maximum temperature (TMPMAX) varies between 16.2 and 37.50 °C, the duration of sunshine (DRINSQ) during the day is between 0 and 13.30 h, the recorded precipitation (RRQUOT) varies between 0 and 19.30 mm during the day, atmospheric pressure (PRESTN) during the day at the station is between 998 and 1017 hpa, relative humidity (HUMREL) varies between 0 and 100%, wind speed (FFVM) is null in the morning and up to 7 m/s in the evening, the most dominant wind direction (DDVMDEG) during the day generally comes from the southwest, and horizontal and vertical wind speeds (V_x and V_y , respectively) are generally low. As the data studied in this paper contain many correlated variables ($VIF > 5$), nine alternative parametric methods were compared to multiple linear regression (MLR, PCR regression, PLS, SPLS, Continuum Regression, Ridge regression, Lasso, and Biased Power regression) to resolve the multicollinearity problem [22]. We conclude that the Lasso and SPLS models give the best results. For this reason, we compare them in this paper with the nonparametric methods in order to select the most appropriate one.

Table 2. Statistics of measured variables at GCR from 01 April to 30 September in 2013 and 2014.

Variable	Min	Max	Mean	St. Dev	VIF
TMPMAX	16.2	37.5	24.5	3.09	4104.12
TMPMIN	8.20	23.50	18.35	3.02	3888.51
TMPMOY	12.40	29.90	21.45	2.88	14,178.12
RRQUOT	0.00	19.30	0.39	1.98	1.78
DRINSQ	0.00	13.30	9.72	2.79	1.66
HUMREL06h	50.00	100.0	87.42	8.00	2.03
HUMREL12h	34.00	95.00	68.32	8.78	2.15
HUMREL18h	28.00	97.00	75.66	9.66	2.13
PRESTN06h	9997.7	1017.3	1008.2	2.97	16.07
PRESTN12h	997.7	1016.5	1008.9	2.91	46.49
PRESTN18h	999	1016	1008	2.88	18.12
FFVM06h	0.00	4.00	1.55	0.80	1.58
FFVM12h	0.00	6.00	3.58	0.98	3.17
FFVM18h	0.00	7.00	3.46	1.04	2.79
DDVM06degre	0.00	360.0	176.4	117.87	1.64
DDVM12hDEG	0.00	360.0	227.3	141.63	2.65
DDVM18hDEG	0.00	360.0	189.2	152.21	2.77
Vx06	−2.95	3.46	−0.05	1.06	2.71
Vx12	−5.91	3.94	−0.59	1.98	4.48
Vx18	−5.91	4.50	−0.10	1.84	5.21
Vy06	−4.00	4.00	0.08	1.38	1.86
Vy12	−3.06	6.00	2.75	1.39	4.25
Vy18	−5.36	6.00	2.79	1.36	4.50
<i>O₃veilleJahid</i>	10.00	130.0	52.83	25.66	1.08

The data studied above required a preliminary treatment for the imputation of missing data by the k-nearest neighbor's method [46] and the agreement between the different meteorological parameters. Multidimensional study on a complete database using a standardized (scaled and centered data) Principal Component Analysis (PCA) [47] was conducted. The objective is thus to identify strong correlations between the different predictors to avoid multicollinearity problems [22]. The correlation circle of variables allows us to visualize the different correlations that exist between the meteorological variables and the *O₃Jahid* variable. Therefore, about 40% of the variation is explained by the first two eigenvalues together (Figure 3).

**Figure 3.** Correlation circle variables (1–2) factorial plane) according to the contributions of variables to PCs: *O₃Jahid* (dashed blue) considered as a supplementary variable.

We distinguished strong correlations between wind (V_x and V_y) and pressure (PRESTN06, PRESTN12, and PRESTN18) parameters on one side and temperature (TMPMAX, TMPMIN, and TMPMOY) on the other. In this respect, using multiple regression model with all of the variables studied is therefore unstable. Indeed, the best adapted model only considers the most significant independent predictors.

3.2. Internal Validation: Goodness of Fit

The goodness of fit of the compared models on the training data (2013 to 2014 period) is assessed by maximizing (R^2) criteria and by minimizing the RMSE measure. We graphically represent the values of the R^2 (Figure 4A) and RMSE (Figure 4B) obtained by adjusting the five models over the training period.

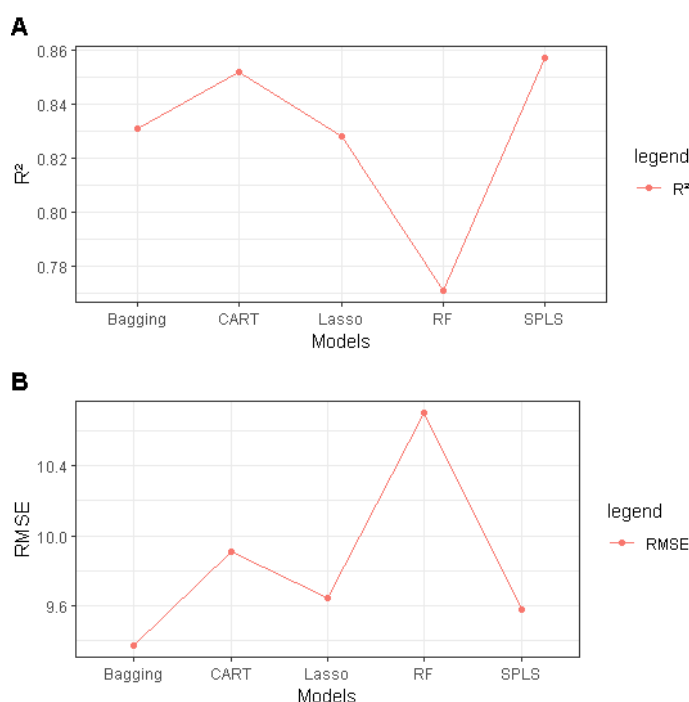


Figure 4. Comparison of models (A): R^2 during summer period of 2013 and 2014; (B): RMSE.

Regarding the goodness of fit of the parametric models, the values of R^2 obtained by SPLS and Lasso correspond to a good fit (0.857 and 0.828, respectively). In the nonparametric models, the CART model shows a good fit of the data ($R^2 = 0.831$), followed by the Bagging model, which gives results similar to the Lasso model. The non-fixed variables retained by the RF model among the initial twenty-four variables provide an explanatory contribution of 79%, which is still not as good compared to the other models. In addition to the R^2 criterion, other performance indices give more precision in terms of the fit quality of models such as the RMSE error (Figure 4B) to select the best adjusted model. At first view, the Bagging nonparametric model outperforms the other models by obtaining the lowest RMSE value (9.342). The other models are very similar in terms of fit data, recording an RMSE error that varies between 9.5 and 9.6, with the exception of the RF model, which has a slightly higher error (9.91). As for the parametric models, SPLS and Lasso register a good fitting in terms of R^2 and RMSE criteria. However, if the objective is to obtain the best predictive model, the RMSE alone is not sufficient to decide on the best forecast model. We therefore use the RMSEP, calculated on the validation period (2015) during the external validation step, to evaluate the predictive quality of each model.

3.3. External Validation: Performance Evaluation

Model performance is evaluated in this phase with the $RMSEP$ and calculated from the 2015's test period, which was not used to adjust the models (Figure 2). Two criteria are calculated:

- $RMSEP_{obs}$, by testing the models on observed meteorological data.
- $RMSEP_{prev}$, by testing the models on the forecasted meteorological data.

Testing the models on observed meteorological data indicates that Lasso and SPLS, respectively, surpass the other methods and remain very similar in terms of predictive capacity. In general, the results obtained from the $RMSEP_{obs}$ range from 11.5 to 14 (Table 3).

The most considered results are those of models testing on real forecasted meteorological data. In terms of predictive capacity, it appears that, among the nonparametric models studied, Bagging gives the lowest value of the $RMSEP_{prev}$ (12.65) followed by its particular case the RF model (12.85) and then the Lasso parametric model (13.00) (Table 3).

However, for operational purposes, it is preferable to identify a simpler model (with few predictive variables) to avoid the problem of a missing forecast rather than a more complex model with a slightly lower $RMSEP$. For this reason, (Figure 5) presents the number of predictive variables retained by each model.

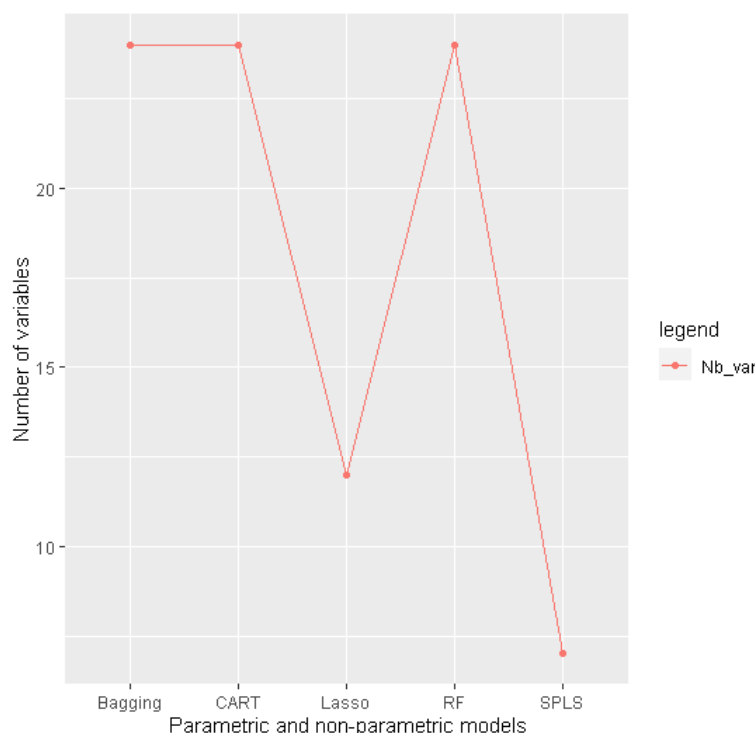


Figure 5. Comparison of model performance according to number of parameters.

This suggests that Lasso and SPLS retain the advantage of a small number of selected predictive variables (12 and 7, respectively) using the parsimonious principle (reducing the number of predictive variables by applying a penalty that leads to fixing the regression coefficients of the non-selected variables to 0). However, the comparison of the performances of each obtained model using the principle of the threefold cross validation technique (Appendix A and Table A2) reveals that the SPLS and Lasso models appear unstable and more sensitive to the modification of the training and test periods of the data. In this regard, the next section proposes a new simpler multiple regression model with fewer predictor variables ensuring a good prediction quality.

3.4. Selected Forecast Model

The principle of this forecast model is to build a step-by-step regression model from the explanatory variables stored by the SPLS and Lasso parametric models. We thus reduce all explanatory variables of a regression from the first obtained results in Section 3.3. The objective is to obtain a forecasting model that is easier to implement and to interpret in order to obtain short-term forecasts. The regression model that is more important and significantly predictors and high predictive capacity was selected. The scheme below presents the construction process of the selected forecast model (Figure 6).

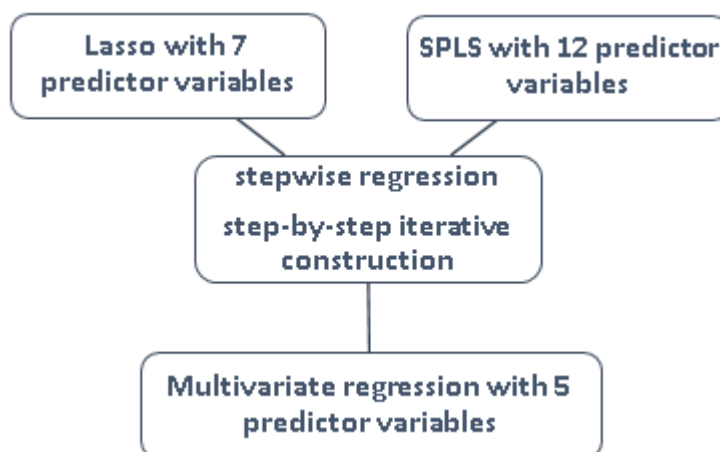


Figure 6. Building process of the final forecast model.

Following the adjustment of the new model on the training data set (2013–2014) (internal validation) and its test for the year 2015 (external validation), the results obtained from the different evaluation criteria by comparing them with other models (Section 3.3) are summarized in Table 3.

Table 3. A Comparison of the final model with other models according to $RMSE$, R^2 , $RMSEP$, and Nb Variables criteria.

Models/Criteria	R^2	$RMSE$	$RMSEP_{obs}$	$RMSEP_{prev}$	Nb Variables
SPLS	0.857	9.576	11.89	13.61	7
Lasso	0.828	9.555	11.58	13.02	12
CART	0.852	9.523	14.16	13.83	24
Bagging	0.831	9.342	12.87	12.65	24
RF	0.771	9.914	13.36	12.85	24
Selected model	0.856	9.60	11.78	12.55	5

The selected parametric model for forecast O_3 daily concentrations retains five significant variables: the sunshine duration ($DRINSQ$), the pressure at 6 h ($PRESTN_{06}$), the horizontal wind at 6 h (V_{x06}), the vertical wind at 6 h (V_{y06}), and the O_3 concentrations from the previous day ($O_{3veilleJahid}$). The model equation is written as Equation (5):

$$O_{3Jahid} = 52.73 + 1.41 * DRINSQ - 1.17 * PRESTN_{06} - 1.33 * V_{x06} + 1.34 * V_{y06} + 23.51 * O_{3veilleJahid}. \quad (5)$$

The results of the statistical tests necessary for the diagnosis and validation of the regression model are statistically significant at the 5% significant level. Table A3 summarizes the test's result.

The regression coefficients obtained from the selected model (Equation (5)) make it possible to explain the impact of the different factors retained on O_3 emissions. Indeed, O_3 concentration is influenced positively by sunshine duration (+1.41), horizontal wind

direction at 6 h (+1.34), and the previous day's O_3 (+23.51) and negatively by pressure at 6 h (−1.17) and vertical wind direction at 6 h (−1.33). These results are similar to those obtained in studies conducted by [3,21,48].

Table 3 compares the results obtained by the selected reduced regression model with those of the five parametric and non parametric models studied in Section 2.2. This comparison is evaluated using the performance indices of the model's adjustment quality on learning data (R^2 and $RMSE$) and prediction quality on test data (2015) ($RMSEP_{obs}$ and $RMSEP_{prev}$). In terms of adjustment, the selected model provides successful results ($R^2 = 0.86$) with an $RMSE$ of 9.60 which remains slightly high compared to the SPLS, Lasso, and CART models. In terms of the quality of prediction on observed test data, the final model is very close to that of the Lasso model with ($RMSEP_{obs} = 11.78$) but remains more accurate than the other models. The validation of the selected model on meteorological forecast data (the summer period of 2015) indicates that it outperforms parametric and nonparametric models in terms of forecasting capacity with the lowest prediction error ($RMSEP_{prev} = 12.55$). Furthermore, the reduced number of regressors used in the final model (Nb variables = 5) has an important advantage compared to other models, namely, (i) easy prediction of O_3 concentrations and (ii) easy interpretation of the prediction model by avoiding data unavailability.

On this subject, the following graph presents a comparison between the observed O_3 concentration (red color) and the forecasted O_3 concentration (green color) using the selected model with five predictors.

Figure 7 shows that O_3 concentration's predictions obtained on forecasted meteorological data in 2015 using the selected model are very close to those observed for the same period. This clearly shows the success of the selected regression model in terms of the accuracy of O_3 forecasts for one-step-ahead (i to $i + 1$ day) in the GCR. This finally chosen model also provides the best performance on indicators related to forecast quality, such as R^2 , $RMSE$, and $RMSEP$ when compared to the results obtained by Lei [18], which compared the MLR and CART models to forecast O_3 concentrations in Macao.

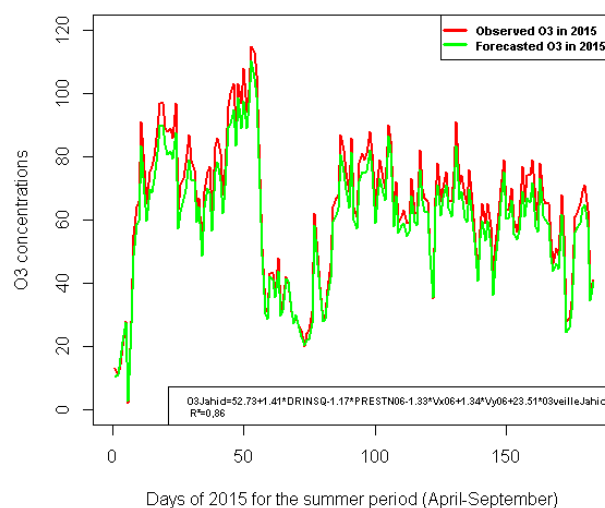


Figure 7. Comparison between observed and forecasted O_3 concentrations.

4. Discussion

In this study, two statistical approaches were compared to predicting daily ozone in the GCR: (i) parametric models and (ii) nonparametric models. The majority of studies performed to forecast daily ozone have used both parametric and nonparametric statistical models in some way to find a significant relationship between meteorological factors, persistence, and O_3 concentrations [3]. The work conducted in this paper is the first of its kind in Morocco, and the results obtained by similar work performed in other countries were discussed and compared. This comparison remains a complex task that must be

carried out with precaution because each country has its own characteristics. In addition, the geographical position, the climate, the history of the data used, and the methodology adopted play very important roles in the development of the forecasting model.

Descriptive statistical analysis of the data; parametric models of the SPLS and Lasso models; and nonparametric models of the CART, Bagging, and RF models were designed to evaluate the best performing model in terms of forecast quality. Several statistical indicators of evaluation were computed, namely R^2 , $RMSE$, and $RMSEP$ (Section 2.3). In terms of the fitting quality, the results of the coefficient of determination R^2 obtained by SPLS and Lasso, alternative models to MLR, were satisfactory ($R^2 = 0.82$, $R^2 = 0.86$) and close to those obtained by the work conducted by [49] (0.6 to 0.90) at different measuring stations, [20] (0.6 to 0.9), [19] (0.60 to 0.80), [50] (0.84). However, Lasso and SPLS appear to be sensitive at the training period level, as they do not maintain their predictive capacity by changing the training period and the test period (Figure 2). Indeed, this instability of the parametric models at the selection level of the complete set of variables is mainly due to the problem of multicollinearity (cf. [22]).

In terms of the nonparametric models, the results obtained for the CART and RF models (0.85 and 0.77, respectively) are similar to those obtained by the author [50] (0.86) who compared the MLR model to the CART model as well as [43] (0.78), who compared MLR and RF. In terms of predictive quality, the studied models have been evaluated by the “ $RMSEP_{prev}$ ” indicator, which is calculated by using forecasted meteorological data for the validation year (2015). The Bagging and RF models showed good predictive capacity (12.65 and 12.85, respectively) compared to CART (13.83), Lasso (13.02), and SPLS (13.61). These results are still very satisfactory compared to the study conducted in Macao by [18] ($RMSE = 24.35$), obtained using the CART model. However, Bagging and RF have the disadvantage of being considered “black box” models with many variables that can produce missing forecasts in practice. Indeed, this type of model has complications related to the optimization of the time allocated to the calculation of the parameters necessary for their adjustment as well as to the efficiency of their implementation. As the purpose of the present study is to implement the first short-term ozone prediction model in the GCR, it is difficult to implement a model retaining 24 predictors such as Bagging and RF (24 variables) to routinely predict O_3 concentrations.

Following a profound evaluation of the performance of the parametric and nonparametric models, the final chosen model is based on multivariate regression model using a stepwise method on independent variables retained in Lasso and SPLS models. The final model (Section 3.4) that considers only five of the most significant predictors such as ($DRINSQ$, $PRESTN_{06}$, Vx_{06} , Vy_{06} , and $O_{3veilleJahid}$) was selected. In fact, these meteorological factors are coherent with the results obtained in the literature [3,7,21]. This final selected model has the best predictive ability ($RMSEP = 12.55$) compared to the other models. These results are consistent with the findings of [33], who found that the MLR model is better in terms of prediction than RF. In the same context, ref. [51] also found that MLR outperforms the machine learning models in predicting daily ozone in India. In addition to its predictive quality in terms of accuracy, robustness, and efficiency, the final model provides significant improvements such as (i) being more stability, (ii) having a lower number of predictor variables, (iii) being resistant to missing data and multicollinearity problems, (iv) facilitating implementation and interpretation, and (v) being able to adapt to changing situations (new measurement stations and inclusion of additional data sets). Indeed, the final selected model provides a suitable platform for forecasting, showing a good performance (Figure 7) for a first forecasting model in the GCR.

For future research aiming to improve results in terms of forecasting quality, new data will be added to ensure the stability of the selected model and to extend the forecasting over all cities equipped with measuring stations in Morocco. The model obtained in this study is the most appropriate for predicting the usual O_3 concentrations. The question that arises here would be to determine if this model could predict pollution peaks as well. For this purpose, it is interesting to focus on the predictive quality of extreme values and threshold

exceedances of ozone. The aim is thus to create an alert system in case of exceedance of the limits. Another important work will be to study the other atmospheric pollutants such as PM_{10} , NO_2 , and SO_2 in Morocco.

Author Contributions: Conceptualization, H.O.; methodology, H.O. and L.B.; software, H.O. and L.B.; validation, L.B. and A.B.; formal analysis, H.O. and L.B.; investigation, H.O., L.B., A.B., and K.K.; resources, H.O., K.K.; data curation, H.O., K.K.; writing—original draft preparation, H.O.; writing—review and editing, H.O., L.B., A.B., and K.K.; visualization, H.O., L.B., A.B., and K.K.; supervision, L.B. and A.B.; project administration, H.O., L.B., and A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this article are confidential. They are provided by the General Directorate of Meteorology in Morocco.

Acknowledgments: This research was supported in part by the CNRS project MAiROC. The General Directorate of Meteorology of Morocco is gratefully acknowledged for providing the necessary data to conduct the present study.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GCR	Grand Casablanca Region
GDM	General Directorate of Meteorology
ALADIN	Limited Area Dynamic Adaptation International Development
GADM	Global Administrative Areas
MLR	Multiple Linear Regression
SPLS	Sparse Partial Least Square
LASSO	Least Absolute Shrinkage and Selection Operator
CART	Classification and Regression Tree
RF	Random Forest
VIF	Variance Inflation Factor

Appendix A. Parametric Models

Appendix A.1. Multiple Linear Regression Model (MLR Model)

In order to describe the parametric models studied in this paper, it is essential to provide the matrix notation of the multiple linear regression model (MLR model) Equation (A1):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad (\text{A1})$$

where $\mathbf{y}$ is a $(n \times 1)$ vector (centred dependant variable), $\mathbf{X}$ is a $(n \times p)$ matrix (matrix of standardized predictors), $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of unknown regression coefficients, and $\mathbf{e}$ is a $(n \times 1)$ vector of random errors. The distribution of $\mathbf{e}$ is assumed to be normal with mean equal to 0 and a variance covariance matrix equal to $\sigma^2 \mathbf{I}$, where $\mathbf{I}$ is the identity matrix [35]. Usually, the variables in $\mathbf{X}$ are centered by subtracting their means and scaled by dividing by their standard deviations. In MLR model, the estimated regression coefficients are defined by

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

In the case of this study, the expression of Equation (A1) becomes Equation (A2):

$$(O_3)_i = \beta_0 + \beta_1(O_3)_{i-1} + \sum_{j=2}^p \beta_j \text{meteovar}_i^j + e_i, \quad (\text{A2})$$

where $(O_3)_i$ is the concentration in O_3 day i , $(O_3)_{i-1}$ is the previous day's O_3 concentration (i.e., persistence), and meteovar_i^j is the value of meteorological variable j day i . For more information about MLR model using in this case, you can refer to [22].

Appendix A.2. Sparse PLS Regression (SPLS)

The Sparse PLS method described in [25] is a direct adaptation of the classic PLS regression method. It is a parsimonious approach that integrates a selection of variables into the PLS by using penalties on the weight norm to make this selection. PLS regression yields latent components $\mathbf{T}_{(m)}$ such that $\mathbf{T}_{(m)} = \mathbf{X}\mathbf{W}_{(m)}$, where m is the number of latent components $w_j; j = 1, \dots, m$. In SPLS regression, w the first weight vector as an optimal solution to Equation (A3):

$$\max_{w \in \mathbb{R}^n} (w^T \mathbf{M} w) \text{ subject to } w^T w = 1 \text{ and } \|w\|_1 \leq \eta, \quad (\text{A3})$$

where $\mathbf{M} = \mathbf{X}^T \mathbf{y} \mathbf{y}^T \mathbf{X}$, $\|w\|_1 = \sum_{j=1}^m |w_j|$ is the L^1 -norm of vector w and $\eta > 0$ is a scalar which controls the degree of sparsity.

The estimated regression coefficients of $\hat{\beta}_{SPLS}$ are calculated by fixing the coefficients of the non-selected variables to 0 and by obtaining the coefficients of the selected variables with the “standard” PLS regression (see Equation (A4)).

$$(\hat{\beta}_{SPLS})_j = \begin{cases} (\hat{\beta}_{PLS})_j, & \text{if } w_j \neq 0; j = 1, \dots, m \\ 0, & \text{if else.} \end{cases} \quad (\text{A4})$$

The interest of the SPLS is twofold: on the one hand, it reduces the number of explanatory variables, which allows for an easier interpretation of the model, and on the other hand, it avoids the problem of multicollinearity by using the PLS regression model. Moreover, SPLS is computationally efficient with a tunable sparsity parameter to select the important variables.

Appendix A.3. Lasso Regression

The Least Absolute Shrinkage and Selection Operator, or Lasso [26,27], is another penalized regression where the L^2 penalty of ridge regression is replaced by an L^1 penalty. This is a subtle change that has important consequences. Indeed, this constraint entails that some of the regression coefficients are shrunk exactly to zero. This means that this regression strategy operates a selection of variables since the unimportant variables are discarded, with their regression coefficients being equal to zero. Formally, the lasso estimator is given as a solution to the following optimization problem designed by Equation (A5):

$$\hat{\beta}_{Lasso} = \underset{\beta \in \mathbb{R}^n, \|\beta\|_1 \leq \delta}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\beta\|^2, \quad (\text{A5})$$

where $\delta \geq 0$.

The parameter δ controls the degree of sparsity: the smaller this parameter, the larger the number of discarded variables. Contrariwise, if δ is larger than $\delta_0 = \|\hat{\beta}_{OLS}\|_1$, then $\hat{\beta}_{Lasso} = \hat{\beta}_{OLS}$. Lasso regression has the double effect of shrinking the β_j coefficients, allowing us to decrease the variance of the regression coefficients as with Ridge regression and, more importantly, to operate an automatic selection of the variables by cancelling out some β_j coefficients.

Appendix B. Nonparametric Models

The nonparametric models used to predict tropospheric ozone in this study are Regression Tree, Bagging, and Random Forests.

Appendix B.1. Classification and Regression Tree (CART)

Ref. [28] is a nonparametric supervised classification or regression method depending on the nature of the variable to be explained (qualitative or quantitative). It is complementary to the above linear regression method in the regression version. It is a method frequently used in air quality prediction [18,26]. This learning method is based on the implementation of a decision tree as a prediction model. The construction of this tree uses the data set for which the value of the target variable is known and then projects the results to the test data. Since the data studied in our paper are quantitative, we are interested in the regression CART tree.

Binary tree's construction: CART is a binary recursive partitioning technique consisting of splitting the data into two groups, resulting in a binary tree, in which the terminal nodes represent distinct classes or categories of data. Cutting is carried out according to simple rules on the explanatory variables, determining the optimal rule that allows us to build two populations more differentiated in terms of values of the variable to be explained. It builds a partition visualized using a binary tree [26]. A classic stopping rule is not to cut out nodes that contain less than a certain number of observations. Terminal nodes, which are no longer cut out, are called the sheets of the tree.

The objective of a decision tree is to divide the forecast space into K separate class such as, $C_1, \dots, C_K$. In the case of the quantitative response variable y , regression tree model can be written as the following Equation (A6):

$$y_i = \sum_{k=1}^{n_f} 1_{C_k}(x_i^1, \dots, x_i^p) \beta_k + e_i, \quad (\text{A6})$$

where n_f is the unknown number of sheets of the tree underlying the method, C_k corresponds to the different classes, and β_k are the coefficients to be estimated by the average in each class of the explained variable.

Pruning: After the first step of building the CART tree, the second step called "pruning" consists in removing a posteriori branches of the tree considered unnecessary to find the best tree-pruned for the maximum tree. The idea is that the maximum tree corresponds to a parsimonious model with a very high variance and a low bias.

The CART method is adapted to the case of many explanatory variables of different types, which allows explicit results to be given with simple decision rules that can be easily interpreted and do not disturb extreme observations (e.g., cases of pollution peaks).

However, one of its inconveniences is that the CART method may not reach the optimal model or it may also present instability in the trees obtained (a light perturbation). Indeed, other methods or practices exist that can provide solutions to this type of problem; the main random aggregation methods of forecasting are namely Bagging and Random Forests.

Appendix B.2. Bagging

Bootstrap Aggregation (or Bagging for short) is a random method firstly introduced by Breiman in [29]. It applies the Bootstrap or Bootstrapping principle [52] to the aggregation of predictors. It allows to average the forecasts of several independent models and to reduce the variance as well as the forecast error. To briefly introduce the Bagging, we consider $f(x)$ a model function of $x = \{x^1, \dots, x^p\} \in \mathbb{R}^p$ and $z = \{(x_1, y_1), \dots, (x_n, y_n)\}$ a sample of distribution F . Considering B independent samples noted $\{z_b\}_{b=1, \dots, B}$, a model aggregation forecast is defined below by Equation (A7) in the case where the variable to be explained is quantitative:

$$\hat{f}_B(\cdot) = \frac{1}{B} \sum_{b=1}^B \hat{f}_{z_b}(\cdot). \quad (\text{A7})$$

The samples so obtained are obviously not independent, but the instability of the trees can make the set completely efficient: each tree has a low bias while its mean is of low variance. The parameters number of trees can be determined in practice by the cross-validation technique [41].

The advantage of this method is the simplicity and facility of adaptation to the implementation of the different modelling methods. However, bagging requires a longer computation time to run the model and the model results are not viewed as a decision tree.

Appendix B.3. Random Forests (RF)

The RF method was presented in [30] and more recently for example in [53]. Its objective is to improve Bagging for binary trees (CART) by introducing “randomization”, which makes it possible to develop the trees in the aggregation more independent by adding chance in the choice of variables used in the model. As its name suggests, a RF consists of aggregating trees of discrimination or regression.

We consider $\mathcal{L}_n$ the Learning data set, $\{\hat{h}(\cdot, \theta_1), \dots, \hat{h}(\cdot, \theta_B)\}$ a collection of predictors per tree, where $(\theta_1, \dots, \theta_B)$ is a series of B independent and identically distributed random variables. The predictor of random forests is obtained by aggregating this collection of trees. A RF is therefore no more than an aggregation of trees dependent on random variables. For example, bagging trees (building trees on bootstrap samples) defines a random forest. A family of random forests is different from others, particularly in terms of the quality of its performance on many data sets. The term “Random Forests” is derived from the fact that individual predictors are, in this case, explicitly predictors per tree and from the fact that each tree depends on an additional random variable (i.e., in addition to the learning data set $\mathcal{L}_n$). A random forest is a part of the family of group methods based on a decision tree, which makes it possible to aggregate a collection of random trees. Every “tree” in the “forest” is trained with a different learning data set. Each of these sets is a subsample of the set of learning data, randomly selected using the Bagging method. The importance of each variable for the forecast can therefore be estimated posteriori. It is possible to use variables that do not have a statistical link with the response; their importance in the forecast is reduced by the model [41].

The algorithm principle of the RF method is to add a random draw of m among the p explanatory variables in the application of Bagging to binary decision trees. The number of explanatory variables m can be calculated by default in the case of regression problems using the formula $m = p/3$. In practice, the parameters number of trees built B and number of variables m require optimization through the application of a cross-validation technique. In general, the RF method avoids over-learning with better performance than decision trees. The parameters are easy to calibrate but its training remains slower and difficult to interpret.

Appendix C. Comparison Table of Parametric and Nonparametric Models

Appendix C.1. Training Period: (2014–2015) (Case 1)

Table A1. Comparison table of parametric and nonparametric models according to the R^2 , $RMSE$, $RMSEPobs$, $RMSEPprev$, and Nb Variables criteria.

Models/Criteria	R^2	$RMSE$	$RMSEPobs$	$RMSEPprev$	Nb Variables
SPLS	0.827	9.299	12.77	13.98	8
Lasso	0.801	9.233	12.77	13.63	14
CART	0.803	9.909	13.69	13.81	24
Bagging	0.802	8.704	13.44	14.04	24
RF	0.774	9.579	16.80	17.40	24

Appendix C.2. Training Period: (2013–2015) (Case 2)

Table A2. Comparison table of parametric and nonparametric models according to the R^2 , RMSE, RMSEPobs, RMSEPprev, and Nb Variables criteria.

Models/Criteria	R^2	RMSE	RMSEPobs	RMSEPprev	Nb Variables
SPLS	0.761	11.40	9.595	8.92	6
Lasso	0.751	10.94	9.151	7.63	13
CART	0.757	11.48	12.41	12.62	24
Bagging	0.734	10.30	10.39	10.19	24
RF	0.701	11.78	10.21	8.40	24

Appendix D

Table A3. Tests results for the final forecast model's diagnostic check.

Test	H_0	p -Value
Normality (Shapiro–Wilk normality)	Residuals normality	0.049
Homoscedasticity (Studentized Breusch–Pagan)	Homoscedasticity	0.8361
Autocorrelation (D–W Autocorrelation)	$\rho = 0$	0.314
Linearity (Harvey–Collier)	Nonlinear relation	0.004

References

- Liu, J.C.; Peng, R.D. Health effect of mixtures of ozone, nitrogen dioxide, and fine particulates in 85 US counties. *Air Qual Atmos Health* **2018**, *11*, 311–324. [CrossRef]
- Lin, X.; Yuan, Z.; Yang, L.; Luo, H.; Li, W. Impact of extreme meteorological events on ozone 346in the pearl river delta, China. *Aerosol Air Qual. Res.* **2019**. [CrossRef]
- Wang, T.; Xue, L.; Brimblecombe, P.; Fat Lam, Y.; Li, L.; Zhang, L. Ozone pollution in China: A review of concentrations, meteorological influences, chemical precursors, and effects. *Sci. Total Environ.* **2017**, *575*, 1582–15961. [CrossRef] [PubMed]
- Khoms, K.; Chelhaoui, Y.; Alilou, S.; Sour, R.; Najmi, H.; Souhaili, Z. Concurrent heatwaves and extreme Ozone (O3) episodes: Combined atmospheric patterns and impact on human health. *Earth Space Sci. Open Arch.* **2020**, *16*, 2020. [CrossRef]
- World Population Prospects United Nations Population Estimates and Projections of Major Urban Agglomerations. (2019 Revision). Available online: <https://worldpopulationreview.com/world-cities/casablanca-population> (accessed on 15 April 2021).
- Yang, L.; Xie, D.; Yuan, Z.; Huang, Z.; Wu, H.; Han, J.; Liu, L. Quantification of regional ozone pollution characteristics and its temporal evolution: Insights from the identification of the impacts of meteorological conditions and emissions. *Atmosphere* **2021**, *12*, 279. [CrossRef]
- Fang, C.; Wang, L.; Wang, J. Analysis of the Spatial–Temporal Variation of the Surface Ozone Concentration and Its Associated Meteorological Factors in Changchun. *Environments* **2019**, *6*, 46. [CrossRef]
- Anenberg, S.C.; Horowitz, L.W.; Tong, D.Q.; West, J.J. An estimate of the global burden of anthropogenic ozone and fine particulate matter on premature human mortality using atmospheric modeling. *Environ. Health Perspect.* **2010**, *118*, 1189–1195. [CrossRef]
- Green, R.; Broadwin, R.; Malig, B.; Basu, R.; Gold, E.B.; Qi, L.; Sternfeld, B.; Bromberger, J.T.; Greendale, G.A.; Kravitz, H.M.; et al. *Epidemiology* **2016**, *27*, 211–220. [CrossRef]
- Van Eijkeren, J.C.; Freijer, J.I.; Van Bree, L. A model for the effect of health of repeated exposure to ozone. *Environ. Model Softw.* **2002**, *17*, 553–562.
- Leelossy, Á.; Molnár, F.; Izsák, F.; Havasi, Á.; Lagzi, I.; Mészáros, R. Dispersion modeling of air pollutants in the atmosphere: A review. *Cent. Eur. J. Geosci.* **2014**, *6*, 257–278. [CrossRef]
- Zhang, J.; Ding, W. Prediction of Air Pollutants Concentration Based on an Extreme Learning Machine: The Case of Hong Kong. *Int. J. Environ. Res. Public Health* **2017**, *14*, 114. [CrossRef] [PubMed]
- Thompson, M.L.; Reynolds, J.; Cox, L.; Guttorp, P.; Sampson, P.D. A review of statistical methods for the meteorological adjustment of tropospheric ozone. *Atmos. Environ.* **2001**, *35*, 617–630. [CrossRef]
- Sousa, S.I.V.; Martins, F.G.; Alvim-Ferraz, M.C.M.; Pereira, M.C. Multiple Linear Regression and Artificial Neural Networks Based on Principal Components to Predict Ozone Concentrations. *Environ. Modell. Softw.* **2007**, *22*, 97–103. [CrossRef]
- Zhang, Y.; Bocquet, M.; Mallet, V.; Seigneur, C.; Baklanov, A. Real-time air quality forecasting, part I: History, techniques, and current status. *Atmos. Environ.* **2012**, *60*, 632–655. [CrossRef]
- Ben Ishak, A.; Ben Daoud, M.; Trabelsi, A. Ozone Concentration Forecasting Using Statistical Learning Approaches. *J. Mater. Environ. Sci.* **2017**, *8*, 4532–4543. [CrossRef]

17. Zhan, Y.; Luo, Y.; Deng, X.; Grieneisen, M.L.; Zhang, M.; Di, B. Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment. *Environ. Pollut.* **2018**, *233*, 464–473. [CrossRef] [PubMed]
18. Lei, M.T.; Monjardino, J.; Mendes, L.; Gonçalves, D.; Ferreira, F. Macao air quality forecast using statistical methods. *Air Qual. Atmos. Health* **2019**, *12*, 1049–1057. [CrossRef]
19. Jahn, S.; Hertig, E. Statistical modelling of combined ozone-temperature events in Europe. In Proceedings of the EGU General Assembly 2020, Online, 4–8 May 2020; EGU2020-1314. [CrossRef]
20. Allu, S.K.; Srinivasan, S.; Maddala, R.K.; Reddy, A.; Anupoju, G.R. Seasonal ground level ozone prediction using multiple linear regression (MLR) model. *Model. Earth Syst. Environ.* **2020**, *6*, 1981–1989. [CrossRef]
21. Iglesias-Gonzalez, S.; Huertas-Bolanos, M.E.; Hernandez-Paniagua, I.Y.; Mendoza, A. Explicit Modeling of Meteorological Explanatory Variables in Short-Term Forecasting of Maximum Ozone Concentrations via a Multiple Regression Time Series Framework. *Atmosphere* **2020**, *11*, 1304. [CrossRef]
22. Oufdou, H.; Bellanger, L.; Bergam, A.; El Ghaziri, A.; Khomsi, K.; Qannari, E. Comparison of Different Regularized and Shrinkage Regression Methods to Predict Daily Tropospheric Ozone Concentration in the Grand Casablanca Area. *Adv. Pure Math.* **2018**, *8*, 793–812. [CrossRef]
23. Bai, L.; Wang, J.; MaID, X.; Lu, H. Air Pollution Forecasts: An Overview. *Int. J. Environ. Res. Public Health* **2018**, *15*, 780. [CrossRef] [PubMed]
24. World Urbanization Prospects—United Nations Population Estimates and Projections of Major Urban Agglomerations. Available online: <https://worldpopulationreview.com/world-cities/casablanca-population> (accessed on 15 April 2021).
25. Wold, H. Estimation of Principal Components and Related Models by Iterative Least Squares. In *Multivariate Analysis*; Krishnaiah, P.R., Ed.; Academic Press: New York, NY, USA, 1966; pp. 391–420.
26. Hastie, T.; Tibshirani, R.; Friedman, J. *The Element of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: Berlin, Germany, 2009; ISBN 978-0-387-84858-7.
27. Tibshirani, R. Regression shrinkage and selection via the lasso: A retrospective. *J. R. Stat. Soc. Ser. B Stat. Method.* **2011**, *73*, 273–282. [CrossRef]
28. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Chapman & Hall: New York, NY, USA, 1984.
29. Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
30. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
31. Souza, A.; Aristones, F.; Pavão, H.; Fernandes, W. Development of a Short-Term Ozone Prediction Tool in Campo Grande-MS-Brazil Area Based on Meteorological Variables. *Open J. Air Pollut.* **2014**, *3*, 42–51. [CrossRef]
32. Robeson, S.M.; Steyn, D.G. Evaluation and comparison of statistical forecast models for daily maximum ozone concentrations. *Almos. Environ.* **1990**, *246*, 303–312. [CrossRef]
33. Li, H.; Zhu, Y.; Zhao, Y.; Chen, T.; Jiang, Y.; Shan, Y.; Liu, Y.; Mu, J.; Yin, X.; Wu, D.; et al. Évaluation de la performance des capteurs de qualité de l’air à faible coût dans une station de haute montagne avec des conditions météorologiques complexes. *Atmosphere* **2020**, *11*, 212. [CrossRef]
34. Kovac-Andric, E.; Brana, J.; Gvozdic, V. Impact of Meteorological Factors on Ozone Concentrations Modelled by Time Series Analysis and Multivariate Statistical Methods. *Ecological Informatics* **2009**, *4*, 117–122. [CrossRef]
35. Chaloulakou, A.; Assimacopoulos, D.; Lekkas, T. Forecasting Daily Maximum Ozone Concentrations in the Athens Basin. *Environ. Monit. Assess.* **1999**, *56*, 97–112. [CrossRef]
36. Di Carlo, P.; Pitari, G.; Mancini, E.; Gentile, S.; Pichelli, E.; Visconti, G. Evolution of Surface Ozone in Central Italy Based on Observations and Statistical Model. *J. Geophys. Res. D* **2007**, *112*, 10316. [CrossRef]
37. Barrero, M.A.; Grimalt, J.O.; Canton, L. Prediction of Daily Ozone Concentration Maxima in the Urban Atmosphere. *Chemom. Intell. Lab. Syst.* **2006**, *80*, 67–76. [CrossRef]
38. Marzuki, I.; Al-Mahfoodh, N.; Samsuri, A.M. Development of Ozone Prediction Model in Urban Area. *Int. J. Innov. Technol. Explor. Eng.* **2019**, *8*. [CrossRef]
39. Scheifinger, H.; Stohl, A.; Kromp-Kolb, H.; Spangl, W. A statistical method for predicting daily maximum ozone concentrations. *Gefahrstoffe, Reinhaltung der Luft* **1996**, *56*, 133–137.
40. Ryan, W.F. Forecasting severe ozone episodes in the Baltimore metropolitan area. *Atmos. Environ.* **1995**, *29*, 2387–2399. [CrossRef]
41. Genuer, R.; Poggi, J.M.; Tuleau, C. Variable selection using random forests. *Pattern Recognit. Lett. Elsevier* **2010**, *31*, 2225–2236. [CrossRef]
42. Gómez-Losada, A.; Asencio-Cortés, G.; Martínez-Álvarez, F.; Riquelme, J.C. A novel approach to forecast urban surface-level ozone considering heterogeneous locations and limited information. *Environ. Modell. Softw.* **2018**, *110*, 52–61. [CrossRef]
43. Stafoggia, M.; Johansson, C.; Glantz, P.; Renzi, M.; Shtein, A.; de Hoogh, K.; Kloog, I.; Davoli, M.; Michelozzi, P.; Bellander, T. A Random Forest Approach to Estimate Daily Particulate Matter, Nitrogen Dioxide, and Ozone at Fine Spatial Resolution in Sweden. *Atmosphere* **2020**, *11*, 239. [CrossRef]
44. Geisser, S. The predictive sample reuse method with applications. *J. Am. Statist. Assoc.* **1975**, *70*, 320–328. [CrossRef]
45. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: New York, NY, USA, 2013.
46. Beretta, L.; Santaniello, A. Nearest neighbor imputation algorithms: A critical evaluation. *BMC Med. Inf. Decis. Mak.* **2016**, *16*, 74. [CrossRef]

-
47. Jolliffe, I.T. *Principal Component Analysis*, 2nd ed.; Springer: New York, NY, USA, 2002.
 48. Alvim-Ferraz, M.C.; Sousa, S.I.; Pereira, M.C.; Martins, F.G. Contribution of anthropogenic pollutants to the increase of tropospheric ozone levels in the Oporto Metropolitan Area, Portugal since the 19th century. *Environ. Pollut.* **2006**, *140*, 516–524. [[CrossRef](#)]
 49. Bekesiene, S.; Meidute-Kavaliauskiene, I.; Vasiliauskiene, V. Accurate Prediction of Concentration Changes in Ozone as an Air Pollutant by Multiple Linear Regression and Artificial Neural Networks. *Mathematics* **2021**, *9*, 356. [[CrossRef](#)]
 50. Lei, M.T.; Monjardino, J.; Mendes, L.; Gonçalves, D.; Ferreira, F. Statistical Forecast of Pollution Episodes in Macao during National Holiday and COVID-19. *Int. J. Environ. Res. Public Health* **2020**, *17*, 5124. [[CrossRef](#)] [[PubMed](#)]
 51. Pandya, S.; Ghayvat, H.; Sur, A.; Awais, M.; Kotecha, K.; Saxena, S.; Jassal, N.; Pingale, G. Pollution Weather Prediction System: Smart Outdoor Pollution Monitoring and Prediction for Healthy Breathing and Living. *Sensors* **2020**, *20*, 5448. [[CrossRef](#)] [[PubMed](#)]
 52. Altman, N.; Krzywinski, M. Ensemble methods: Bagging and random forests. *Nat. Methods* **2017**, *14*, 933–934. [[CrossRef](#)]
 53. Cutler, A.; Cutler, R.; Stevens, J.R. *Random Forests*. Chapter 5: *Ensemble Machine Learning: Methods and Applications*; Springer: Berlin/Heidelberg, Germany, 2012. [[CrossRef](#)]