Evaluation and Compaison of Different Daily Ozone Statistical Prediction Models for the Grand-Casablanca Area

Halima Oufdou, Lise Bellanger, Amal Bergam, Kenza Khomsi

J. Adv. Math. Stud.
Vol. **11**(2018), No. 2, 376-390
http://journal.fairpartners.ro

EVALUATION AND COMPARISON OF DIFFERENT DAILY OZONE STATISTICAL PREDICTION MODELS FOR THE GRAND-CASABLANCA AREA

HALIMA OUFDOU, LISE BELLANGER, AMAL BERGAM AND KENZA KHOMSI

Abstract. This work deals with the forecasting of daily tropospheric ozone episodes in the Grand-Casablanca area. We present a comparison of different statistical predictive models, derived from various methods. We fit them on observed data collected and validated by the National Direction of Meteorology of Morocco. Finally, we compare them in order to deliver recommendations on the real-time forecasting model to be adopted routinely to plan the daily ozone in the Grand-Casablanca area.

1. INTRODUCTION

Air pollution means the contamination of the environment by a chemical, physical or biological agent that blurs the natural characteristics of the atmosphere. Today, air pollution has become the main environmental health risk in the world (WHO, 2015) due to high concentration of human activities. In the recent years, Morocco has experienced significant urban, industrial and demographic development and may therefore be highly impacted and sufferings from the degradation of its environment, mainly air quality. This promotes risks of acute respiratory, chronic and cardiovascular diseases. Air quality studies are still rarely undertaken in Morocco and that's why this work deals with this issue in the framework of a scientific research aiming forecasting daily tropospheric ozone (O3) in the Grand-Casablanca area using a statistical models. The city of Casablanca is known as:

- The economic capital and the largest city containing more than (insert the number of inhabitant);
- An important city containing many industrial units and activities;
- A city with an important traffic mainly during the rush hours;
- Containing an important air quality measurement network with an important air quality database.

It is rather difficult to forecast air pollution resulting from complex phenomena at different scales of time and space.For the purpose of this study, a statistical approach that consists in determining a statistical relationship between the response variable (O3) and predictors from either meteorological measurements or pollution measurements is used. Also, a short-term forecast (day for the next day) on a small scale is included. The main objective is to develop a first comparison of two statistical models adjusted to predict daily average concentration

Paper presented to MOCASIM 2017.
Key words and phrases: Tropospheric ozone, applied statistics, forecast, linear and not linear models, methods of regression, comparison of statistical models.

of ozone measured in the Grand-Casablanca area. This paper is organized as follows. In Section 2, we describe the data and the two different methods. In Section 3 ,we present the obtained by both models. In Section 4, we discuss the result obtained on our data.

2. MATERIALS AND METHODS

Air quality in the Grand-Casablanca is monitored by the National Direction of Meteorology using a network of background fixed stations for monitoring pollutants concentrations.

2.1. **Materials.** For the purpose of this study, a comprehensive database provided by the DMN from 01/01/2013 to 31/12/2015, was used. The following data were considered:

• Daily observed and already controlled meteorological data:(minimum, maximum and average temperature measured in °C at 56 m heights, wind direction in units of degrees, wind velocity (unit: m/s) sunshine duration (unit: hour), humidity (unit: percentage), precipitation (unit: mm)) (see Appendix).

• Air pollution data: ozone (O3), nitrogen dioxide (NO2), Sulphurdioxide (SO2) and particulates matters (PM10) measured at Jahid's station, one of four fixed measures of background station in the network of the DMN (Sidi Othman, Jahid, Mohammadia-Prefecture and Mohammadia-Khansae) installed in the Grand-Casablanca region (Fig. 1). The following notations are used with the common measurement unit ($\mu g/m^3$) (see Appendix).

FIGURE 1. Geographical situation of four fixed stations in the Grand-Casablanca area

In this study, we focused on ozone (O3) as a secondary pollutant of photochemical origin that is found in the lower layers of the atmosphere called tropospheric ozone (O3). It is formed from reactions of other pollutants in the atmosphere between nitrogen oxides and hydrocarbons, in presence of solar radiation. The tropospheric ozone stays for a long period of time in the atmosphere and increases in the summer, due to the presence of solar radiation. Results and graphs of this study are obtained using the statistical software R.

2.2. **Methods.**

2.2.1. *Data processing.* The available data require an important preliminary preprocessing step in order to choose the station characterized by:

- Geographical situation near the road traffic.
- The longest series of data, with few missing values. Indeed, we are interested here in the background "Jahid" fixed measuring station to build our statistical forecasting models (Fig. 1 and Fig. 3).

Before this, that's necessary to identify missing values, occur when no data value is stored for the variable in an observation. Missing values are a common occurrence and can have a significant effect on the final results. That's why we need to choose the appropriate imputation method for replacing these values, Fig. 3.

K-nearest neighbors (KNN) is a simple algorithm that stores all available missing values and classifies new values based on a similarity measure by a distance function [1]. We briefly present its simple algorithm: Make

- $Y = y_{ij} \in \Re^{n*p}$: The rectangular matrix of the data for $Y_1, \ldots, Y_p$, p variables and n observations.
- $(y_{ij})_{\text{miss}} = Y_{i*j}$: The missing values of the data for p variables and n observations.

(1) Choice of an integer k: $1 \geq k \geq n$.
(2) Calculate distanced (y_i^*, y_i) between observed value y_i and missing value y_i^*, $i = 1, \ldots, n$.
(3) Hold the k observations nearest of missing value $y_{(i_1)}, \ldots, y_{(i_k)}$ for which these distances are smaller to calculate their average.
(4) Allocate to the missing values the average of the values of the neighboring [2],

$$(y_{ij})_{\text{miss}} = y_{i*j*} = \frac{1}{k}(Y_{(i_1)} + \cdots + Y_{(i_k)}).$$

The schema following present the principle of these method (Fig. 2).

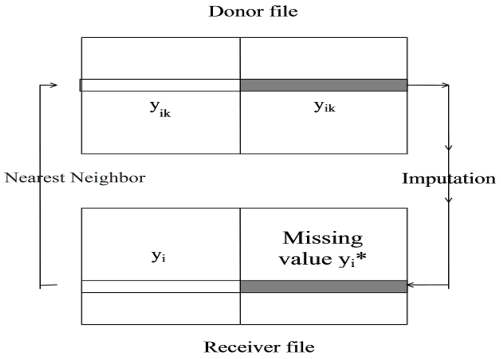

FIGURE 2. Principle of K nearest neighbor's method

After imputation by KNN, we acquire a exhaustive database for the summer period from 2013 to 2014 with 366 days and 22 quantitative variables including the predictantozone. The summer of the year 2015 will be used as test data to test the precision of the predictions of the models.

In addition, to complete this data processing, we explore our data using multivariate data analysis and more particularly Principal Components Analysis (PCA). PCA is a classical statistical exploratory tool to make a synthesis of the relations between variables orobservations in a dataset visualized as a set of coordinates in a high dimensional space. It reveals the internal structure of the data in a way that best explains the variance in the data. PCA give a lower-dimensional picture of the data set, projection of this object when viewed from its most informative viewpoint.

Most often the projection is done in a space of dimension 2 called the first factorial plane. The classical graphical representations are: (i) Representations of individuals and (ii) Representations of quantitative variables (correlation circle). PCA is realized on complete data after imputation and on standardized variables constituted of 366 days and 22 variables. It allows us to summarize the relation between explanatory variables and detect those strongly correlated together which will cause multicollinearity problems and consequently, parameters instability in the models. As our study, all variables are quantitative, we present in this part two statistical forecasting models adapted. The first model is the simple and classical multiple linear regression model and the second one corresponds to the Classification And Regression Tree (CART) analysis [3]. These two models have been adjusted on the training data set completed after simple imputation consisting of summer2013 and 2014 and validated on testing data set for summer 2015 to evaluate models performances according to classical indicators of the quality of model fitting.

2.2.2. *Multiple linear regression.* The method of multiple linear regression allows to explain a single quantitative response variable y according to several variables X^j $(j = 1, \ldots, p)$ [4] in the following form:

$$M: \quad y = X\beta + \varepsilon$$

or

$$y_i = \beta_0 + \beta_1 X_i^1 + \cdots + \beta_p X_i^p + \varepsilon_i, \quad i = 1, \ldots, n,$$

where: y_i is the response variable to be explained; $X^1, \ldots, X^p$ are the available with explanatory variables; $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is the vector of unknown parameters; $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ is the term of errors i.i.d. (independent and identically distributed), ε_i following a normal distribution $N(0, \sigma^2)$, where σ^2 is the variance and $i = 1, \ldots, n$ day.

The classical regression model is based on the following assumptions on the errors:

- Homoscedasticity (constant variance) of the error scan be checked by (Breusch-Pagan test) [5]: $\text{var}(\varepsilon_i) = \sigma^2$, $\forall i = 1, \ldots, n$.
- Statistical independence of the errors (in particular, no correlation between consecutive errors) can be checked by (Durbin-Watson test) [6]: $\text{cor}(\varepsilon_i; \varepsilon_{i'}) = 0$, $\forall i \neq i'$.
- Normality of the error distribution: errors $\varepsilon_i \sim N(0, \sigma^2)$ can be checked by (Shapiro-Wilks test) [7].

The internal validation phase consists in verifying these assumptions. Multicollinearity could be defined as a high degree of linear dependency among several independent variables.Within the framework of regression model, it is also necessary to suppose that X is full ranked to make Ordinary Least Squares (OLS) possible. Indeed, a perfect multicollinearity violates this assumption making OLS impossible. When a model is not full ranked, the inverse of X cannot be defined and there can be an infinite number of least squares solutions. However, existence of multicollinearity does not violate the OLS assumption but have consequences on estimated variances that could be inflated. To detect multicollinearity, we can compute correlation coefficients of independent variables and also use PCA because

high correlation coefficients do not necessarily imply multicollinearity. We can quantify the severity of multicollinearity by checking related statistics, such as variance inflation factor (VIF) for $\widehat{\beta}_j$ according to the following formula: $\text{VIF}(\widehat{\beta}_j) = \dfrac{1}{1-R_j^2}$, with R_j^2 is the coefficient of determination of the regression model where X^j is the response variable adjusted on the $(p-1)$ other variables. It reflects all other factors that influence the uncertainty in the coefficient estimates. There is no formal criterion for determining the bottom line of VIF a rule of thumb is that a VIF greater than 10 roughly indicates significant multicollinearity. It's important to avoid groups of variables with high VIF to obtain a stable model.

For selecting the important and significant variables which constitute the reduced regression model for the dataset, AIC estimates the quality of each model among several others regression models. It would apply a consistent approach to selecting which significant variables [8].

The Akaike information criterion (AIC) defined by: $\text{AIC} = n * \log\left(\dfrac{\text{SCR}}{n}\right) + 2(l+1)$, with n the number of observations, SSE the Sum of Squared Error of the estimated model and $l \leq p$ the number of explanatory variables retained.

2.2.3. *Binary trees of decision (or* CART *(Classification and Regression Tree)).* CART [9] is a nonparametric supervised classification or regression method depending on the variable to be explained is qualitative or quantitative. It is complementary to the above linear regression method. It is a binary recursive partitioning technique consisting in splitting the data into two groups, resulting in a binary tree, whose 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 which allows 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 [10]. A classification and regression tree is constructed iteratively, by cutting at every step the population into two subsets according to the test that produces a minimum of residual variance. The construction of the tree stops when the variance decreases. The second phase called pruning is an alternative way to build a decision tree model uses a large tree first and then prune it to optimal size by removing nodes that provide less additional information [11].

2.2.4. *Models comparison.* To assess the accuracy of these models on our data, we used a cross validation technique that allows to assess how the results of a statistical analysis will generalize to an independent data set. In our study, we partionne the available data into two complementary data sets: (i) summers 2013 and 2014 (called training set) using to adjust the models and (ii) summer of 2015 (called the validation set or testing set) using to "test" the models in the training phase. The models are fitted on the training set then, the fitted model is used to predict the ozone responses for the observations of the validation set. The performance of the models is measured using some indicators to compare statistical models [12] according validation phase. In phase of internal validation, we can use following criteria to verify the quality of model adjustment:

- The multiple correlation coefficient R^2 for linear model and CART,

$$R^2 = \left[\sum_{i=1}^{n} (\widehat{y}_i - \overline{y}_i)^2\right] \Big/ \left[\sum_{i=1}^{n} (y_i - \overline{y}_i)^2\right]$$

allows to compare also the quality of the adjustment;

• The adjusted multiple correlation coefficient R^2 adjusted (R_j^2) indicates how well terms fit a curve or line, but adjusts for the number of terms in a model $R_j^2 = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right]$. k is the number of independent regressors, i.e. the number of variables in the model, excluding the constant. In external validation, these criteria are used to verify quality of the model forecasting [13].

• The Mean Absolute Error (MAE) measures the average magnitude of the errors in a set of predictions, without considering their direction, $\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y}_i|$, where y_i is the actual value of y for day i, $\widehat{y}_i$ the y-value for object i predicted with the model under evaluation n is the number of days for which $\widehat{y}_i$ is obtained by prediction.

• The Root Mean Squared Error of Prediction (RMSEP) differences between prediction and actual observation, $\text{RMSEP} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}$, more this criterion is small, more variance of the prediction error is low and reduced means error.

3. RESULTS

3.1. **Data processing.**

3.1.1. *Choice of the measuring station.* To choose the background station the most representative we compare distributions of ozone concentrations measured by four stations of measurements obtained by presenting them in the form of boxes with a mustache on a single graph (Fig. 3).

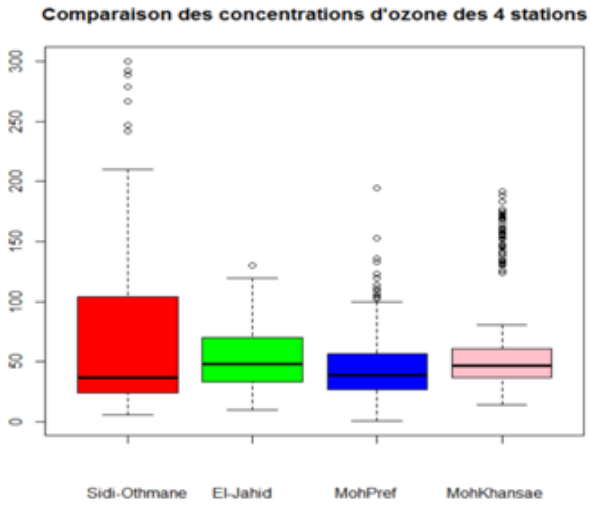

FIGURE 3. Box-plot of the four measurement stations

The Box-plots present a comparison of ozone concentrations measured by 4 background fixed stations located in the Grand-Casablanca with initial data prior to imputation. In this study, we interest in the measuring station 'Jahid' recommended by the DMN given the low number of missing values in its recorded data set compared to other stations (see Fig. 3) and its downtown located nearto road traffic. Other stations that will be analyzed later behave

differently according to their location. That explains the visible elevated concentrations on the chart (Fig. 3).

3.1.2. *Diagnosis of missing values.* Identification and treatment of missing data is an important step in the data preparation phase in order to impute them to build a complete summer database.

• First of all, it may be of help to get an overview of the data set, e.g. of the proportion of missing values. It may be even more interesting to analyze if there are certain combinations of variables with missing values. The following graphs (Fig. 4) obtained with VIM R library presents: Left plot: Bar plots for the proportion of missing values in each variable. Right plot: Aggregation plot showing all combinations of missing (grey) and non-missing (yellow) parts in the observations.

The left plot shows that the variables FFVM06h, FFVM12h, FFVM18h, Vx06, Vx12, Vx18, Vy06, Vy12 and Vy18 (see Appendix) contain only 1% missing data while the highest amount is rather in the other four variables. The O3 concentrations variables present the following proportions according to the corresponding measuring station:

- Station El Jahid (6%),
- Station SidiOthmane (9%),
- Station Mohammedia Prefecture (11%),
- Station Mohammedia Khansae (18%).

EL Jahid station contains the lowest proportion of missing data on summer 2013-2015.

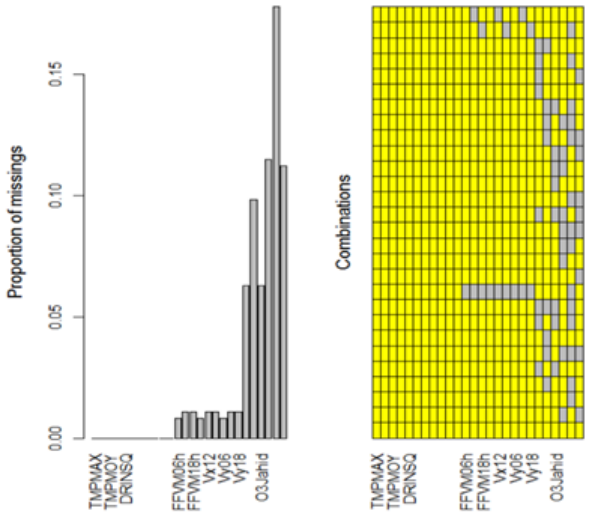

FIGURE 4. Identification of missing data

3.1.3. *Imputation of missing data.* We use in this case K-nearest neighbor procedure, that is a simple type hot-deck imputation method [14]. We choose here the distance between K equal to 10 nearest observations.

3.1.4. *Principal Components Analysis (*PCA*).* The PCA is conducted on complete data and standardized variables (centered and reduced). We present in Fig. 5, all explanatory variables (resp. all the days of the year) are projected in correlation circle (resp. individual's

scatter plot) with O3Jahid as supplementary variable which summarize 40% of the total inertia(the inertia is the total variance of data set). We add supplementary variables "O3Jahid" (blue) to help to interpret the dimensions of variability and because these variable will be our predictant in the next step.

- Correlation circle

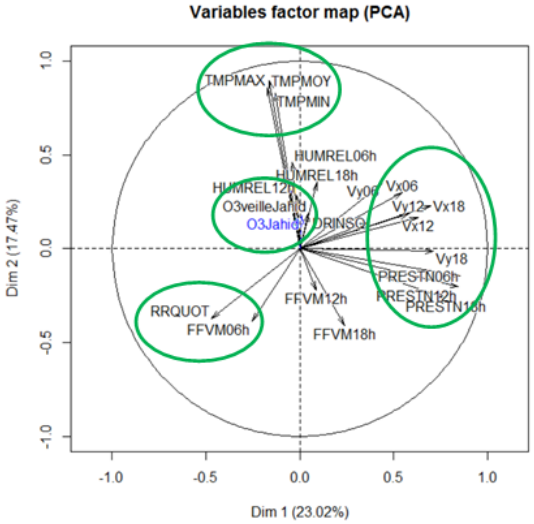

FIGURE 5. Representation of variables on the first plane

We note that arrows Vx06, Vx12 defined variables, Vy18 and Vx18 are well presented on the circle and positively correlated with respect to the first axis. TMPMIN, TMPMAX and TMPMOY variables are well represented with respect to the second main axis. The remaining variables are not well presented on the first factorial plane, but they may be explain in the other dimensions.

- Individuals and variables scatter plot

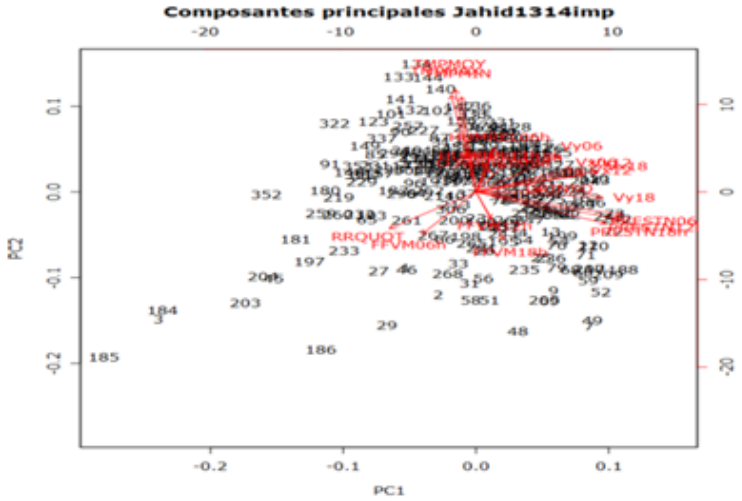

FIGURE 6. Individuals and variables scatter plot on the first plane

The analysis of factorial plan allows to observe the days of the year according weather conditions presented in correlation circle. Thus, it is possible to make groups of days having similar weather features (e.g. the days explained by high temperatures or wind etc.) (Fig. 6).

This first analysis allowed us to clean up the initial data set and then to better understand the relationships between the variables. The existence of problem of multicollinearity is no doubt, what in a modeling framework can be problematic. We present now the built statistical models on summer period of 2013 to 2014 (training period) and to validate it on summer 2015 (external validation).

3.2. **Multiple linear regression model.**

3.2.1. *Complete model.* The complete model containing 24 explanatory quantitative predictors and have a R^2 of 0.8579 and a R^2_{aj} of 0.8479 showing that the quality of Adjustment is relatively good. However, we note that the temperature parameter is not significant while theoretically, these variables should be highly correlated with O3Jahid [15]. That is why, it's necessary to assess the multicollinearity. We present below the variance inflation factor (VIF) values corresponding to all variables (Table 1).

Table 1. Calculation of variable's VIF on complete model

TMPMAX	TMPMIN	TMPMOY	RRQUOT	DRINSQ
4081.59	3869.10	14104.02	1.77	1.65
HUMREL06h	HUMREL12	HUMREL18h	PRESTN06h	PRESTN12h
2.03	2.14	2.13	15.94	45.94
PRESTN18h	FFVM06h	FFVM12h	FFVM18h	Vx06
17.82	1.516	2.89	2.56	2.04
Vx12	Vx18	Vy06	Vy12	Vy18
2.61	2.57	1.82	3.63	3.80
O3veilleJahi				
1.09				

We note that TMPMAX, TMPMIN, TMPMOY, PRESTN06h, PRESTN12h and PRESTN18h factors have a very high VIF value (VIF $>$ 10) which indicates significant multicollinearity problem. The complete regression model obtained is not adapted because it contain several variables not significant and of multicollinearity problem. We decided to construct a reduced model taking into account a subset of significant explanatory variables among those available and having the largest coefficient of determination possible. Then we assess the internal validation of the reduced model.

3.2.2. *Reduced model.* We obtain a reduced model from the complete model using a stepwise variables selection procedure according to Akaike Information Criterion AIC [8]. The summary of the model is presented in Table 2.

We obtained a reduced model with 4 significant regressors (TMPMIN, TMPMAX, DRINSQ and O3veilleJahid), with a R^2_j of $89,7\%$ that indicates a good adjustment. However, we observed that the estimated coefficients of parameters TMPMIN and TMPMAX are opposed while they are positively correlated. The DRINSQ and O3veilleJahid are positively correlated with O3. The internal validation of this reduced model requires verification of the hypothesis mentioned in the Methods section.

Results are presented in Fig. 7.

TABLE 2. Summary of the linear reduced regression model

Residuals					
Min	1Q	Median	3Q	Max	
-20.78	-5.608	-0.77	5.20	20.40	
Coefficients					
	Estimate	Std.Error	T value	Pr(>|t|)	
Intercept	3.59	3.69	0.97	0.33	
TMPMAX	-0.90	0.24	-3.65	0.0002	***
TMPMIN	0.90	0.25	3.65	0.0003	***
DRINSQ	0.63	0.17	3.68	0.0003	***
O3veilleJahid	0.92	0.02	53.27	< 2e-16	***
Signif. code :	0 '***'	0.001 '**'	0.01 '*'	0.05 '.' 0.1	' ' 1
Residual Standard Error : 8.053 on 345 degrees Of freedom					
Multiple R-squared: 0.8971, Adjusted R- squared: 0.896					
F-statistic: 752.2 on 4 and 345 DF, p-value: < 2.2 e-16					

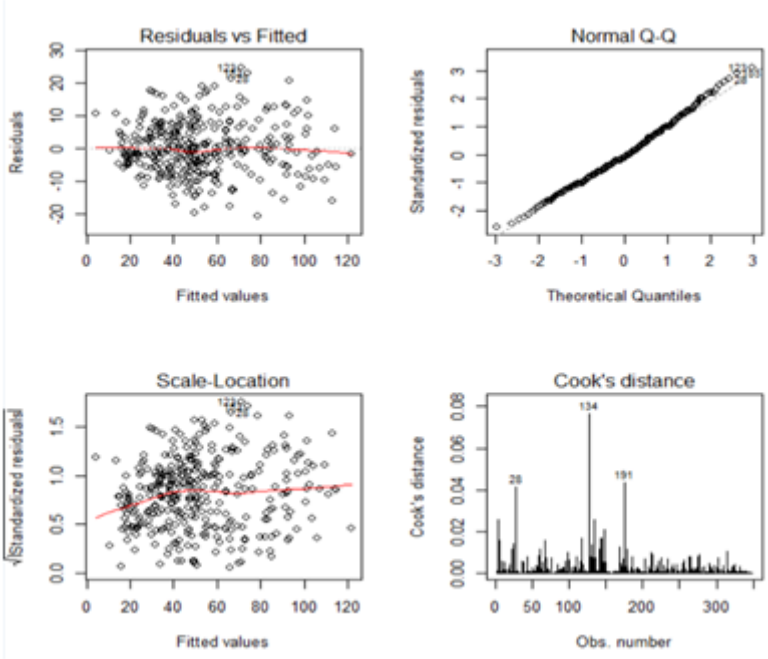

FIGURE 7. Graphical internal validation of the reduced model

By referring to graphs and results of residuals hypothesis tests associated to internal validation of the reduced model, we note that all p-values are inferior at 5% which explain the test satisfaction. The calculus of the VIF gives the following results:

TABLE 3. Calculation of variable's VIF on reduced model

Variable	TMPMIN	TMPMAX	DRINSQ	O3veilleJahid
VIF	2.94	2.91	1.19	1.04

The VIF of TMPMIN and TMPMAX is higher than VIF of DRINSQ and O3veilleJahid but it remained smaller than 10 which show that TMPMIN and TMPMAX are moderately

correlated. If we delete one of these two variables, it becomes not significant, that's why it's preferable to retain the reduce model at 4 regressors. this reduced regression model possess internal validity and can be now compared to CART model.

3.3. **Model CART.** Full binary decision tree obtained by the CART method is built through important variables which explain ozone's concentrations and define his prediction.

• Complete CART. Figure 8 illustrates a simple decision tree model that includes a target variable O3Jahid according to 24 continuous explanatory variables. The main components of a decision tree model are nodes and branches and the most important steps in building a model are splitting, stopping, and pruning. The application of CART model on our data set gets the following regression tree:

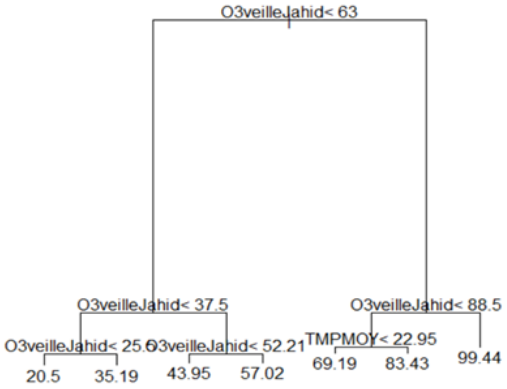

FIGURE 8. Tree CART for the prediction of the O3Jahid variable

The first division corresponds to nodes "2" and "3" created by O3veilleJahid in 63 $\mu g/m^3$ value. We can see at the right of the tree the TMPMOY variable which appears to the value 22.95. We can see in right of tree the TMPMOY variable appears in 22.95 value. We notice that the variable O3veilleJahid is the most active in this full tree(ie the persistence). We then proceed to the step of pruning to reduce the number of explanatory variables and get the optimal tree.

• Reduced CART. We use a pruning procedure to are move the unrepresentative branches (Fig. 9):

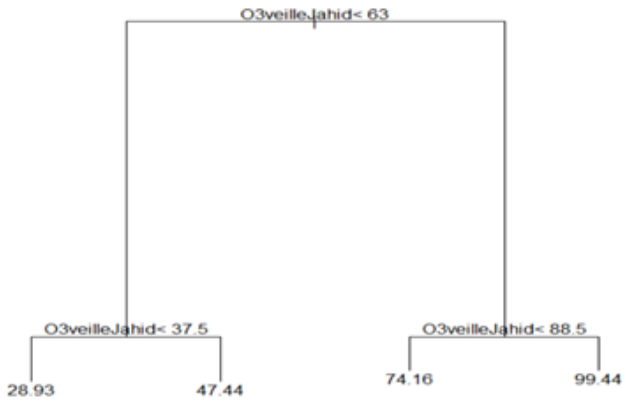

FIGURE 9. Tree CART pruned to predict the O3Jahid variable

We note as well that the concentrations of ozone in the Jahid station are characterized mainly by ozone concentrations of the day before. By considering the tree of regression as a particular model of regression, we can calculate R^2 of the model and check of the hypotheses of internal validation (homoscedasticity, Normality) (Fig. 10).

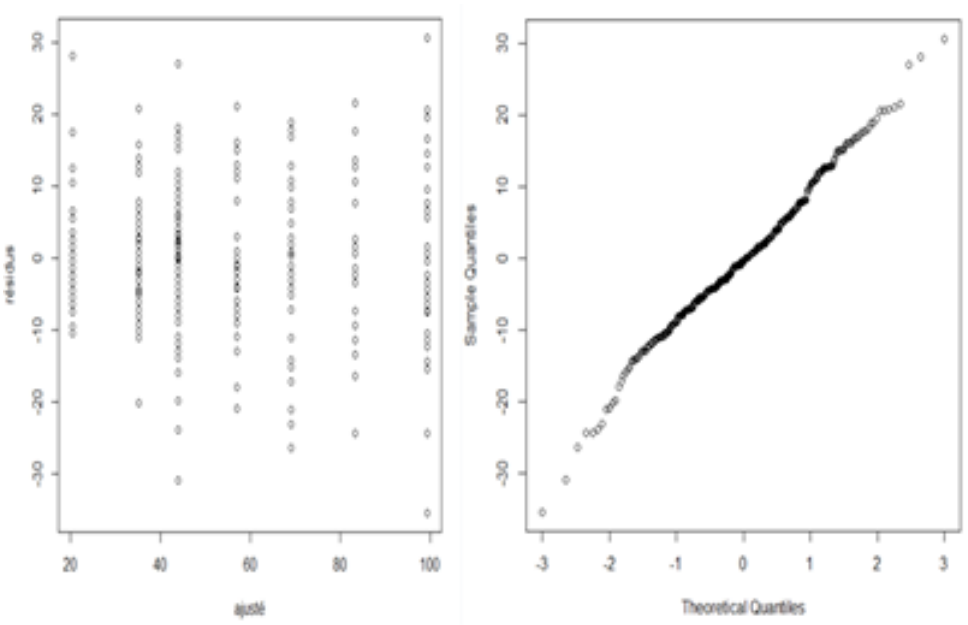

FIGURE 10. Graphical internal validation for CART model

According to the graphs of internal validation of the model CART, we can admit that the supposition of homoscedasticity and normality is satisfied. We also obtain a coefficient of determination R^2 equal to 89%. We can carry out conventional operations of a regression such as forecasting on new data (summer data observed on the year 2015). The results of the external validation (validation test) will be presented in the form of comparison between these two models on the same data set (summer 2015) in the next part.

3.4. **Comparison of two predictive models.** To choose the optimal model for the data of Jahid station in the Grand-Casablanca, we compare the two models using the summer data observed on the validation set (ie summer 2015) according to the criterion of the Root Mean Squared Error of Prediction (RMSEP). The objective of this comparison is to minimize this criterion [16]. The results are presented in the following table:

TABLE 4. Calculation of the RMSEP for the two models

Reduce Linear Model Multiple	CART Model
11.19	12.57

We notice that the RMSEP of the reduced regression model is smaller than that of tree CART, we so conclude that multiple regression to 4 regressors model is most appropriate for our data. We represent the predictions of both models as well as the ozone concentrations observed in the summer period of 2015 on the same chart (Fig. 11).

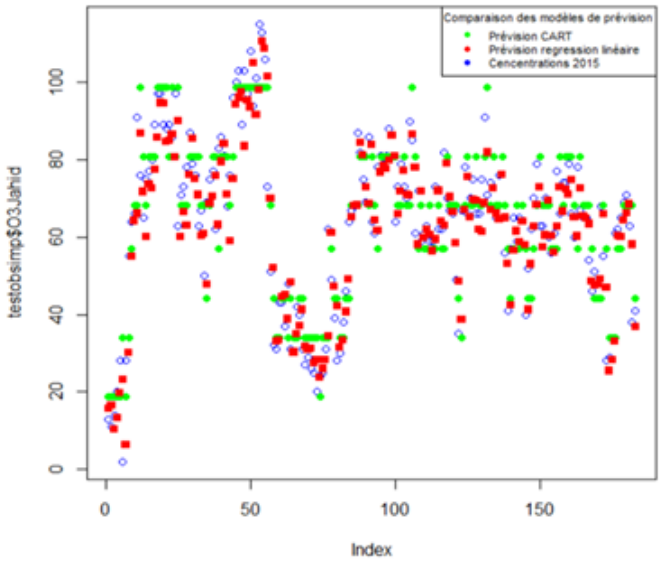

FIGURE 11. Comparison between observed and predicted values of O3 obtained with two models

The comparison chart between the predictions of both studied models and observations shows that the predictions of the regression model are closest to the observed concentrations than the predictions of the CART model.

4. CONCLUSION AND PERSPECTIVES

In summary, we presented in this paper, a description and acomplete analysis data obtained from the DMN in the data preprocessing. Then, we studied two statistical models of forecast of the concentrations in ozone in the region of Grand-Casablanca. The Multiple Linear Regression and CART method have been built on training data (internal validation) and applied on test data (external validation) to compare themfrom the point of view of model fit and prediction.

The results obtained allow to admit at this step that the most appropriate model for the data of Jahid station in Grand-Casablanca area is the model of multiple linear regression to 4 regressors. However it's necessary to ameliorate it to avoid the moderated problems of multicollinearity. Indeed, the existence of a moderate correlation between two regressors of temperature which engender, consequently, the problem of the multi-collinearity.

Choosing linear regression and CART to forecast O3 could be discussed there exists other methods more adapted to prevent multicollinearity problem. It would be interesting to compare the performance predict O3 on Jahid station of a large spectrum of shrinkage regression models:

(i) Continuum regression [17] chosen from a continuum of candidates among which we find methods of analysis related to OLS estimation, Principal Components Regression [18], Partial least squares regression [19],

(ii) penalized regression regrouping Ridge [20] and Lasso [21],

(iii) Biased Power Regression [22].

When the best model is held (after the internal and external validation phases), we shall use it in explanatory variables forecasts so as to obtain a model implementable and who supplied forecasts with routine.

Acknowledgements. The authors thank the Direction de la Metrologie Nationale, Morocco, for providing data.

This work is supported by the Centre National de la Recherche Scientifique (CNRS, PICS MAiROC).

Appendix

Abreviation	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 06h	%
HUMREL12h	Relative humidity at 12h	%
HUMREL18h	Relative humidity at 18h	%
PRESTN06h	Pressure at the station level at 06h	HPA
PRESTN12h	Pressure at the station level at 12h	HPA
PRESTN18h	Pressure at the station level at 18h	HPA
FFVM06h	Wind force at 06h	m/s
FFVM12h	Wind force at 12h	m/s
FFVM18h	Wind force at 18h	m/s
DDVM06h	Wind direction at 06h	Degree
DDVM12h	Wind direction at 12h	Degree
DDVM18h	Wind direction at 18h	Degree
Vx06	Horizontal wind at 06h	m/s
Vx12	Horizontal wind at 12h	m/s
Vx18	Horizontal wind at 18h	m/s
Vy06	Vertical wind at 06h	m/s
Vy12	Vertical wind at 12h	m/s
Vy18	Vertical wind at 18h	m/s
O3veilleJahid	Ozone concentrations of the day before	$\mu g/m^3$
O3veille	Ozone concentrations	$\mu g/m^3$

REFERENCES

[1] L. Bellanger and R. Tomassone: *Exploration de Données et Méthodes Statistiques: Data Analysis & Data Mining avec R. Collection Références Sciences*, Editions Ellipses, Paris, 2014.

[2] https:www.math.univ-toulouse.fr/besse/Wikistat/pdf/st-m-app-idm.pdf

[3] R. Genuer and J.M. Poggi: *Arbres CART et Forêts aléatoires. Importance et sélection de variables*, 20 Jan 2017, arXiv:1610.08203 [stat.me], 45 pages.

[4] P-A. Cornillon and E. Matzner-Lober: *Régression avec R*, Springer Verlag, France, 2011.

[5] T. Breusch and A. Pagan: *A simple test for heteroscedasticity and random coefficient variation*, Econometrica, **47**(1979), No. 5, 1287-1294.

[6] J. Durbin and G. Watson: *Testing for serial correlation in least squares regression. II*, Biometrika, **38**(1951), No. 1/2, 159-179.

[7] S. Shapiro and M. Wilk: *An analysis of variance test for normality (complete samples)*, Biometrika, **52**(1965), No. 3/4, 591-611.

[8] G. Mélard: *Méthodes de Prévision à Court Terme*, Les Éditions de l'Université de Bruxelles, Coédition Ellipses, 2007.

[9] L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone: *Classification and Regression Trees*, Chapman & Hall, New York, 1984.

[10] B. Ghattas: *Prévision des pics d'ozone par arbres simples et agrégés par Bootstrap*, Revue de Statistique Appliquée, **47**(1999), No. 2, 61-80.

[11] Y. Song and Y. Lu: *Decision tree methods: applications for classification and prediction*, Shanghai Archives of Psychiatry, **27**(2015), No. 2, 130-135.

[12] S. Abudu, C-L. Cui, J-P. King and K. Abudukadeer: *Comparison of performance of statistical models in forecasting monthly stream flow of Kizil River, China*, Water Science and Engineering, **3**(2010), No. 3, 269-281.
[13] A. Sayegh, S. Munir and T.M. Habeebullah: *Comparing the performance of statistical models for predicting PM10 concentrations*, Aerosol and Air Quality Research, **14**(2014), No. 3, 653-665.
[14] R.R. Andridge and R. Little: *A review of hot deck imputation for survey non-response*, Int. Stat. Rev., **78**(2010), No. 1, 40-64.
[15] Julian J. Faraway: *Practical Regression and Anova using R*, ©1999, 2000, 2002 Julian J. Farawa.
[16] L. Bel, L. Bellanger, V. Bonneau, G. Ciuperca, D. Dacunha-Castelle, C. Deniau, B. Ghattas, M. Misiti, Y.Misiti, G.ppenheim, J-M. Poggi and R. Tomassone: *Eléments de comparaison de prévisions statistiques des pics d'ozone*, Revue de Statistique Appliqué, **47**(1999), No. 3, 7-25.
[17] R. Sundberg: *Continuum regression and ridge regression*, J. R. Stat. Soc. Ser. B. Stat. Methodol., **55**(1993), No. 3, 653-659.
[18] W. Massy: Principal component regression in exploratory statistical research, J. Amer. Statist. Assoc., **60**(1965), 234-256.
[19] P. Geladi and B.R. Kowalski: *Partial least-squares regression: A Tutorial*, Analytica Chimica Acta, **185**(1986), 1-17.
[20] A. Hoerl and W. Kennard: *Ridge regression: biased estimation for non-orthogonal problems*, Technometrics, **12**(1970), No. 1, 55-67.
[21] R. Tibshirani: *Regression shrinkage and selection via the lasso*, J. R. Stat. Soc. Ser. B. Stat. Methodol., **58**(1996), No. 1, 267-288.
[22] E-M. Qannari, A. El Ghaziri and M. Hanafi: *Biased power regression: a new biased estimation procedure in linear regression*, Electron. J. Appl. Stat. Anal., **10**(2017), No. 1, 160-179.
[23] B. Ghattas, L. Mary, P. Renzi and D. Robin: *The ozone in the French department of Bouches-du-Rhône and a forecasting methodology*, Pollution Atmospherique, **32**(2000), No. 14,413-426.
[24] G.E.A.P.A. Batista and M.C. Monard: *K-Nearest Neighbour as Imputation Method: Experimental Results* (in print), Technical Report, ICMC-USP, 2002.
[25] H. Oufdou, L. Bellanger and A. Bergam: *Comparaison de modèles statistiques pour prévoir l'ozone journalier dans la région du Grand-Casablanca*, Conférence Internationale Francophone Apprentissage artificiel & Fouille de données et 23ème rencontre de la Société Francophone de Classification (AAFD) & (SFC), Marrakech, Maroc, (2016), p. 306.
[26] H. Oufdou, L. Bellanger and A. Bergam: *Comparison of different daily ozone statistical prediction models for the Grand-Casablanca*, in Proc. Second International Conference on Modelling and Scientific Computing in Mathematical Engineering (MOCASIM 2017), FST of Marrakech, (2017), p. 133.
[27] J.-M. Poggi and B. Portier: *PM10 forecasting using clusterwise regression*, Atmospheric Environment, **45**(2011), No. 38, 7005-7014.
[28] www.r-project.org

Laboratoire MAE2D FP Larache
B.P. 745 Poste Principale, Larache, 92004 Morocco
E-mail address: `oufdouhalima@gmail.com`

Laboratoire Jean Leray
2, Rue de la Houssiniere
B.P. 92208, 44322 Nantes Cedex 03, France
E-mail address: `lise.bellanger@univ-nantes.fr`

Laboratoire MAE2D FP Larache
B.P. 745 Poste Principale, Larache, 92004 Morocco
E-mail address: `bergamamal11@gmail.com`

Direction de la Meteorologie Nationale
Casablanca, Morocco
E-mail address: `k.khomsi@gmail.com`