Journal of Materials and Environmental Sciences
ISSN : 2028-2508

Copyright © 2017,
University of Mohammed Premier
Oujda Morocco

J. Mater. Environ. Sci., 2017, Volume 8, Issue 12, Page 4532-4543

http://www.jmaterenvironsci.com/

Ozone Concentration Forecasting Using Statistical Learning Approaches

A. Ben Ishak [1], M. Ben Daoud [2], A. Trabelsi [3]

1,2,3. Université de Tunis, ISGT, LR99ES04 BESTMOD, 2000, Le Bardo, Tunisia

Received 18 Oct 2016,
Revised 22 Feb 2017,
Accepted 27 Feb 2017

Keywords

- Ozone forecasting;
- SVR;
- RF;
- Variable selection;
- Selection bias

A. Ben Ishak
anis_isg@yahoo.fr
+216 97 549 940

Abstract

In this paper, we are interested in the statistical modeling and forecasting of the daily maximum ozone concentration in three monitoring stations from Tunisia. A large number of explicative variables has been considered in our study. We have focused our attention on the problem of variable selection in order to improve the forecasting performance. To achieve our goal, we have used Support Vector Regression (SVR) and Random Forests (RF). The main novelties of this paper are: the variety and originality of the approaches for variable selection in regression, and the audaciousness to deal with a sticky situation characterized by a relatively big pannier of explicative variables compared to the number of observations. The experimental results demonstrate that Random Forests outperform Support Vector Regression in variable ranking and selection. Finally, it was shown that the forecasting accuracy is at least preserved, for the three stations, when using only the selected variables.

1. Introduction

There is a growing interest, in day-to-day, in air quality variation. Especially, Atmospheric pollutants concentration forecasting is evermore an important issue in air quality monitoring.

Naturally, humans are constantly exposed to many dangerous pollutants and it is often hard to know exactly which pollutants are responsible for causing sickness. Indeed, Air pollution is responsible for major health effects and diseases and for increases in mortality rates [1]. However, it is almost impossible to isolate pollutants but we can reduce their harmful effects by modeling and forecasting them in order to take necessary precautions.

1.1. *Causes and effects of ozone concentrations*

Particularly, ground-level ozone (O_3) represents a major air pollution problem, both for public health and for environment. Ozone is a reactive oxidant that forms in trace amounts in two parts of the atmosphere: the stratosphere (the layer between 20-30 km above the earth's surface) and the troposphere (ground-level to 15 km). Stratospheric ozone, also known as "the ozone layer", is formed naturally and shields life on earth from the harmful effects of the sun's ultraviolet radiation. Near the earth's surface, ground-level ozone can be harmful to human health and plant-life and is created in part by pollution from man-made (anthropogenic) and natural (biogenic) sources.

Tropospheric ozone is one of the most preponderant air pollutants in urban areas. It accumulates in or near large metropolitan cities during certain weather conditions and typically exposes tens of millions of people worldwide every week during the summer to unhealthy ozone concentrations [2]. Every summer, ground level ozone concentrations rise and cause episodes of photochemical summer smog. This phenomenon is the cause of well recognized public health distress especially for people suffering from respiratory diseases. An ozone level above some well known threshold causes negative effects on biotic health [3-5]. Indeed, tropospheric ozone is an irritating and reactive gas which rather harmful for human health, and affects other important parts of our daily life such as climate, farming, tourism etc. [6,7]. Moreover, it is responsible for increases in mortality rates during episodes of high concentrations [8].

In light of the health effects of ground-level ozone, an accurate ozone alert forecasting system is necessary to warn the public before the ozone reaches a dangerous level [9]. Moreover, this system can help local authorities to look for short-term management strategies and to encourage people to voluntarily reduce emissions-producing activities in order to avoid bad pollution episodes.

Ground-level ozone depends on a sophisticated chemical and physical process as a function of many known and unknown factors. It has been an active topic for air quality study, an interdisciplinary field among atmospheric research, geochemistry and geophysics. Ozone is not directly emitted by human activities. In troposphere, it is a secondary pollutant which formation depends on a complex cycle [8]. Ozone is produced by atmospheric photochemical reactions that need solar radiation. Its production is lead by volatile organic compounds (which include hydrocarbons) and nitrogen oxides concentrations, both emitted by anthropogenic activities. This harmful pollutant accumulates and scatters owing to three processes:

- In the presence of high temperature and solar radiation, the primary pollutants such as Nitrogen Oxides (NOx) and Volatile Organic Compounds (VOCs) participate in photochemical reactions and contribute to the increase of ozone level.
- Vertical transport of stratospheric air, rich in ozone, into the troposphere [10].
- Horizontal transport due to the wind that brings O3 produced in other regions [11].

1.2. *Related work*

It's very difficult to model ozone concentrations due to the complex interactions between pollutants and meteorological variables [12]. A wide range of statistical approaches was presented in the previous studies to predict O_3 concentrations. On the one hand Ortiz-García et al. [8] used principal component analysis, on the other hand Yi and Prybutok, Spellman, Gómez-Sanchis et al. and Pires et al. [7,13-15] used artificial neural network. In a more recent work, Feng et al. [16] combine neural network with support vector machines.

Machine learning models have shown good performance over a wide range of applications including that related to environmental studies [3,17]. Specifically, support vector regression algorithms (SVR) showed amazing performance on ozone short-term prediction [8]. In the fore-mentioned paper, the authors tried several configurations to obtain the best set of explicative variables to predict O_3 concentration. At each try, neither the variables nor the size of the best set were automatically chosen. In the last stage, they just used statistical tests to verify the significance of incorporating different variables into the model. Genuer et al. [18] provided, in their work mainly methodological, some experimental insights about the behavior of the variable importance index based on random forests. To highlight their methodological insights, they conducted many applications both in classification and regression. Before ending their paper, they considered a benchmark regression ozone dataset from the R package mlbench. Apart from that, to our knowledge, there are few works that have tried to analyze the relevance of input variables in ozone and other pollutants concentration prediction [15,19-21].

1.3. *Purposes and outline of this paper*

In this paper, we conduct a comparative study between the two increasingly used statistical learning methods namely Support Vector Regression and Random Forests (respectively, SVR and RF henceforth). The problem of variable selection within a nonlinear regression framework is investigated in order to improve the forecasting quality. Our variable selection procedure is performed in two steps: once all the variables are ranked in a decreasing order of importance according to the SVR and RF scores, we applied a stepwise forward algorithm in order to retrieve the subset of the most explicative variables [22]. The daily maximum ozone concentration is modeled in three monitoring stations from Tunisia. The Tunisian authorities monitor air pollution by means of the National Network for Monitoring Air Quality (RNSQA). This network contains 15 fixed monitoring stations installed all over in the country. The choice of the studied stations depends on the availability of data provided by the Tunisian National Agency of Environment Protection. The three stations considered here are located at:

- Gabes, located in the southeastern Tunisia and near 406 km from the capital Tunis, is one of the biggest industrial cities in Tunisia. Consequently, it is one of the most polluted regions characterized by the massive presence of industrial sites (such as the Tunisian Chemical Group (GCT)) with elevated environmental impact activities.
- Ghazela, located in the northeastern Tunisia in the northern suburbs of the capital, is a polluted conurbation region characterized by an important vehicular traffic.
- Manouba, located in the northeastern Tunisia in the western suburbs of the capital, is a polluted urban region characterized by the presence of some industries and important vehicular traffic.

The main contribution of this work is threefold: to compare the two increasingly used statistical learning methods for regression, to propose a mixed cooperative procedure using SVM and RF for variable selection and to improve O_3 forecasting by using automatic and statistically efficient variable selection approaches. The rest of the article is organized as follows. After this introduction, Section 2 presents *Data description and pretreatments*, exposes SVR and RF methods and gives the essential on how variable relevance scores are computed. In Section 3, we report and discuss the numerical experiments carried on real-world environment dataset. Finally our study will be closed by conclusions and some possible perspectives.

2. Material and methods

Various panniers of explicative variables have been used in the previous works for the purpose of ozone modeling and forecasting [8,25-27]. This variety depends on the availability of measured variables and the objectives of the study. In our study we do not care neither about the number of used explicative variables nor about their contribution to explain the ozone variation as our main goal is to pick out the best statistically.

2.1. Data description and pretreatments

The dataset used in this study consists of daily maximum ozone concentrations (maxO_3), other pollutants (SO_2, NO_2, NO and PM_{10}) and meteorological data observed in three monitoring stations from Tunisia. The first station is installed at Gabes, the second station is at Ghazela and the third one is localized at Manouba. Each database contains 103 observations from 20/06/2014 to 30/09/2014. As the ozone concentration reaches its peak usually in summer, we have chosen this severe period. The datasets were collected from the Mourouj central station of the National Agency for Environmental Protection (ANPE), which acts under the supervision of the ministry of the environment and sustainable development in Tunisia. All the stations monitoring air quality on Tunisian territory are operating on a continuous basis managed by the RNSQA, under the tutorship of the ANPE. In this work we use thirty-six explicative variables to explain the daily maximum ozone concentration. These variables are grouped into four categories; meteorological indicators, other pollutants, some O_3 hourly concentrations at day *j* and delayed maximum ozone concentrations. Table 1 summarizes all explicative variables.

Table 1: The thirty-six explicative variables

Type	Variable	Definition
Meteorological predictors	Tmax	daily maximum temperature (in °C)
	T:8^h,12^h,17^h	temperature measures for day *j* at 8^h, 12^h and 17^h
	WSmin	daily minimum wind speed (in m/s)
	WSmax	daily maximum wind speed (in m/s)
	DWSmin	wind direction associated to WSmin (discrete:1−8)
	DWSmax	wind direction associated to WSmax (discrete:1−8)
	WDdom	daily dominant wind direction (discrete:1−8)
	RHmin	daily minimum relative humidity (in percentage)
	RHmax	daily maximum relative humidity (in percentage)
	SRmax	daily maximum solar radiation (in W/m^2)
Other pollutants	SO_2min	daily minimum concentration of Sulfur dioxide
	SO_2max	daily maximum concentration of Sulfur dioxide
	SO_2:8^h,12^h,17^h	SO_2 concentration for day *j* at 8^h, 12^h and 17^h
	NO_2min	daily minimum concentration of Nitrogen dioxide
	NO_2max	daily maximum concentration of Nitrogen dioxide
	NO_2:8^h,12^h,17^h	NO_2 concentration for day *j* at 8^h, 12^h and 17^h
	NOmin	daily minimum concentration of Nitric oxide
	NOmax	daily maximum concentration of Nitric oxide
	PM_{10}min	daily minimum concentration of Particulate Matter
	PM_{10}max	daily maximum concentration of Particulate Matter
Ozone concentrations	O_3:8^h,12^h,17^h	Ozone concentration for day *j* at 8^h, 12^h and 17^h
Lagged maxO_3	maxO_3 $(j - t)$	maxO_3 concentration of days $j - t$, $t = 1,..., 7$

In some previous studies, it was shown that the first lagged maxO_3 is an important predictor of its current value at the j^{th} day. In his work, Ghattas [23] was limited only to one lag of maxO_3. Here we consider seven delayed maximum ozone concentrations from $j-1$ to $j-7$. The best predictors to consider in the model will be statistically identified hereafter.

We note that the variables associated with wind direction are transformed from degree to categorical data from 1 to 8. Indeed, the disc is divided into eight equal sectors from north = 1, north-east = 2,. . . , south = 5,. . . , to northwest = 8. This is the wind compass describing the eight principal bearings used habitually in meteorology to categorize wind direction.

Figure 1 shows the boxplots of maxO_3 concentrations (in *ppb*) for each monitoring station. In addition to these boxplots, Table 2 gives a summary of basic statistics to complement the synthetic view.

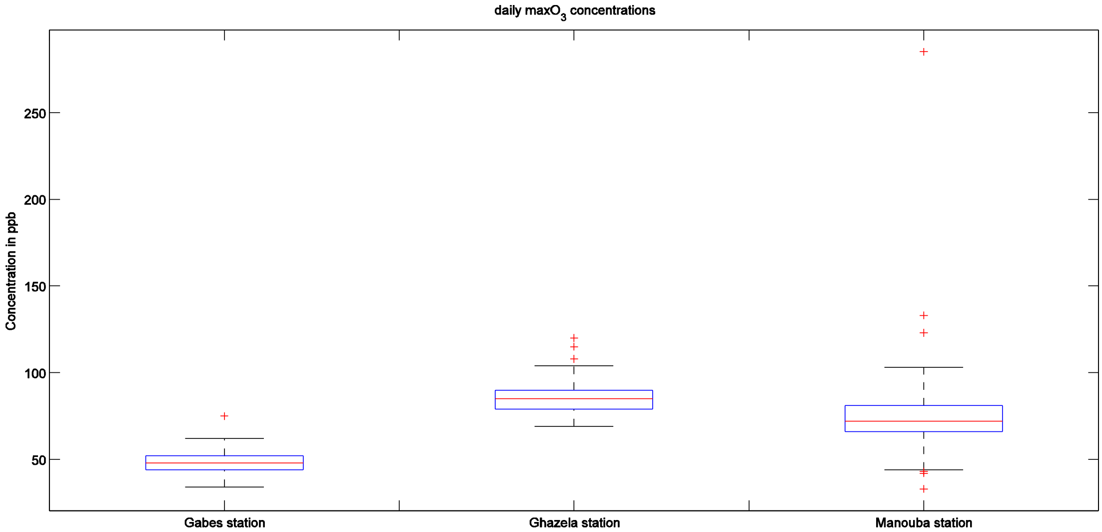

Figure 1: Boxplots of the daily maximum ozone concentrations for Gabes, Ghazela and Manouba stations

Table 2: Descriptive statistics of daily maximum ozone concentrations for the three stations

Statistic/Station	Gabes	Ghazela	Manouba
Minimum	34	69	33
1st Quartile	44	79	66
Median	48	85	72
Mean	48.20	85.38	74.77
3rd Quartile	52	90	81
Maximum	75	120	285
Std. Dev	6.31	8.82	26.41

As it can be seen, the three monitoring stations are different from the maxO_3 concentration distributions. We note a large variability in the maxO_3 values for the three stations. Moreover, we can see that Ghazela and Manouba stations record high levels more often than Gabes station. This result is not surprising given that Ghazela and Manouba are polluted conurbation regions characterized by an important vehicular traffic. Missing data is a ubiquitous problem in evaluating experimental measurements such as related with air quality monitoring. This is due to instrument calibration or malfunction. The treatment of missing values represents an important step in the data mining process. Obviously, we cannot more usual obtain good results from poor or insufficient data. Thus, the three collected raw databases present some missing values. The percentages of missing values are around 1.51%, 1.59% and 6.65% for Gabes, Ghazela and Manouba stations respectively. To handle this problem of missing values, we have used an imputation technique based on a multivariate imputation by chained equations developed by Van Buuren and Groothuis-Oudshoorn [26] and implemented in the MICE algorithm on the software R freely downloadable from http://cran.rproject.org/.

2.2. Statistical tools

In our experiments we will use SVR [27,28] and RF [29] to identify the best explicative variables for the maxO_3 concentrations. Several scores of importance can be derived from the SVR model. According to the results of the intensive comparative study conducted by Ben Ishak [30], we will compare here only the SVR-based score ∂G_α and the *RFS* score derived from RF model.

2.2.1 Variable importance based on SVR

This section presents a description of the basic idea and formulation of SVR and variable importance score based on it. The idea of SVR is to look for a function [30][31]:

$$f(x) = \left(w, \Phi(x)\right)_H + b \quad \text{with } w \in X, b \in \mathbb{R} \tag{1}$$

One way to ensure the flatness of function f is to minimize the Euclidean norm of its weight vector w. We can write this problem as a convex optimization problem which is called the dual problem. This optimization problem can be solved more easily in its dual formulation:

$$maximize \begin{cases} \sum_{i=1}^{n} y_i\,(\widehat{\alpha}_i - \alpha_i) - \varepsilon \sum_{i=1}^{n}(\widehat{\alpha}_i + \alpha_i) \\ -\frac{1}{2}\sum_{i,j=1}^{n}(\widehat{\alpha}_i - \alpha_i)(\widehat{\alpha}_j - \alpha_j)\left(K\left(x_i, x_j\right) + \frac{1}{c}\,\delta_{ij}\right) \end{cases} \tag{2}$$

$$subject\ to \sum_{i=1}^{n}(\widehat{\alpha}_i - \alpha_i) = 0,$$

$$\widehat{\alpha}_i \geq 0, \alpha_i \geq 0, i = 1, \ldots, n$$

where the hyperparameter C is the error cost and it determines the trade-off between the flatness of f and the amount up to which deviations larger than ε are tolerate. ε presents the width of the tube ε-tube. $K\left(x_i, x_j\right) = \langle\varphi(x_i).\varphi(x_j)\rangle$ is the used kernel, $\widehat{\alpha}_i, \alpha_i\ \ for\ i = 1, \ldots, n$ are the Lagrangian multipliers associated to the primal problem constraints and δ_{ij} being the Kronecker symbol.For more details we can consult [30]. The variable selection criterion used in our study have been proposed by Rakotomamonjy as a supplementary criterion for variable ranking because it is relatively cheaper to compute [31]and it was presented as follow:

$$G_\alpha\left(\alpha, \widehat{\alpha}\right) = \sum_{i=1}^{n}(\alpha_i + \widehat{\alpha}_i) \tag{3}$$

[30] showed its performance against other criteria.

2.2.2 Variable importance based on RF

A forest is an ensemble of trees like in real life. In the random forests framework for regression problems, the most widely used score of importance of a given variable, suggested by Breiman [29], is the increasing in Mean Squared Error (the "MSE") when permuting at random the observed values of this variable in the Out-Of-Bag samples (the "OOB"). The accuracy of a random forest's prediction can be estimated from these OOB data as:

$$OOB_{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{\hat{y}}_{iOOB})^2 \tag{4}$$

where $\bar{\hat{y}}_{iOOB}$ denotes the average prediction for the i^{th} observation from all trees for which this observation has been OOB.

The RF importance score for the j^{th} variable is determined as follows:

– For each tree t = 1 , . . . , ntree in the forest we compute the OOB Mean Squared Error as the average of the squared deviations of OOB responses from their respective predictions:

$$OBB_{MSE}^{t} = \frac{1}{|OBB^t|}\sum_{i \in OBB^t}(y_i, \hat{y}_{i,t})^2 \tag{5}$$

where OBB^t contains data not included in the bootstrap sample used to construct t, OBB^t denotes its cardinality and $\hat{y}_{i,t}$ indicates the prediction for the i^{th} observation from tree t.

– For each variable j = 1 , . . . , p we compute the OOB Mean Squared Error of each tree t =1 , . . . , ntree on the associated perturbed OOB sample, $\widetilde{OOB}^{t\ j}$ by randomly permuting the values of the j^{th} variable:

$$\widetilde{OBB}_{MSE}^{t}{}^{j} = \frac{1}{|\widetilde{OOB}^{t\ j}|}\sum_{i \in \widetilde{OOB}^{t\ j}}(y_i, \hat{y}_{i,t}) \tag{6}$$

– For each variable j in each tree t the following difference is calculated:

$\widetilde{OBB}_{MSE}^{t}{}^{j}$

Finally, the RF importance score of variable j is obtained as the average over all ntree trees of the previous differences:

$$RFS_j = \frac{1}{ntree}\sum_{t=1}^{ntree}\left(\widetilde{OBB}_{MSE}^{t}{}^{j} - OBB_{MSE}^{t}\right) \tag{7}$$

2.2.3 Variable selection procedure

The procedure of variable selection is performed in two steps: once all the variables are ranked in a decreasing order of importance according to ∂G_α and the *RFS*, we applied a stepwise forward algorithm in order to retrieve the subset of the most explicative variables. The stepwise forward strategy, firstly introduced in the work of Ghattas and Ben Ishak [22], based on a sequential introduction of variables. A sequence of nested increasing models $M^{(k)}$, k = 1 , 2 , . . . , p, is constructed invoking at the beginning the k most important variables, by step of 1. When p is huge therefore k becomes too large, the additional variables are invoked by blocks. Then the error rate of each model $M^{(k)}$ is estimated by stratified random splitting. The set of variables leading to the model of smallest error rate is selected. Unlike the search-space procedures hereinabove mentioned, this algorithm allows to automatically identifying the size of the selected subset of pertinent predictors.

2.2.4 Performance measures

To evaluate and to compare the forecasting effectiveness of the different models, we have adopted several statistical performance metrics. In addition to the classical metrics, various new types of metrics were discussed and were deeply compared in the literature [34-36]. Overall, it can be stated that none of the efficiency metrics performs ideally. Each of them has specific pros and cons which have to be taken into account during model evaluation. However, some measures can be more complementary and allow together to make fair evaluation. The statistical metrics considered here were successfully used in climatic, hydrologic, and environmental domains, and especially, in previous studies of PM_{10} and other air pollutants [17,20,37]. The selected metrics that will be used are: the Root Mean Squared Error (*RMSE*), the Mean Absolute Error (*MAE*), The Mean Absolute Percent Error (*MAPE*), the factor of 2 (FA_2) and the factor of 1.25 ($FA_{1.25}$), the refined index of agreement (d_r), and finally the coefficient of efficiency (E_1). It is important to emphasize that the significances of these statistical metrics are not equal, but they complete themselves strongly. Their formulas are expressed as follows:

$$RMSE = \sqrt{\frac{1}{l}\sum_{i=1}^{l}(O_i - P_i)^2}\ , \tag{8}$$

$$MAE = \frac{1}{l}\sum_{i=1}^{l}|O_i - P_i|\ , \tag{9}$$

$$MAPE = \frac{1}{l}\sum_{i=1}^{l}\left|\frac{O_i - P_i}{O_i}\right| \times 100\%\ , \tag{10}$$

$$FA_2 = \frac{1}{l}\sum_{i=1}^{l}\chi_{[0.5,2]}\left(\frac{P_i}{O_i}\right), \tag{11}$$

$$FA_{1.25} = \frac{1}{l}\sum_{i=1}^{l}\chi_{[0.8,1.25]}\left(\frac{P_i}{O_i}\right), \tag{12}$$

$$d_r = \begin{cases} 1 - \frac{\sum_{i=1}^{l}|O_i - P_i|}{2\sum_{i=1}^{l}|O_i - \bar{\boldsymbol{O}}|}, if\ \sum_{i=1}^{l}|O_i - P_i| \le 2\sum_{i=1}^{l}|O_i - \bar{\boldsymbol{O}}| \\ \frac{\sum_{i=1}^{l}|O_i - P_i|}{2\sum_{i=1}^{l}|O_i - \bar{\boldsymbol{O}}|} - 1, otherwise \end{cases}, \tag{13}$$

$$E_1 = 1 - \frac{\sum_{i=1}^{l}|O_i - P_i|}{\sum_{i=1}^{l}|O_i - \bar{\boldsymbol{O}}|}, \tag{14}$$

where O_i and P_i are the observed and the predicted values, respectively, $\bar{\boldsymbol{O}}$ is the mean of the observed values, and $\chi_I(x)$ is the indicator function which equals 1 if $x \in I$ and 0 otherwise. In general, good predictive models are associated with simultaneous achievement of small values for *RMSE*, *MAE* and *MAPE*. The other metrics serve to reinforce the judgment. The FA_2 and $FA_{1.25}$ factors provide the proportion of cases for which the values of the ratios $\frac{P_i}{O_i}$ fall in the range [0.5,2] and [0.8,1.25], respectively. The d_r statistical index of model performance is bounded by -1 and 1, and it measures similarity between the modeled and the observed tendency. In general, it is more rationally related to model accuracy than are other existing indices [35]. Finally, to date, the E_1 coefficient is the main competitor with d_r [34]. For the last four metrics, the higher the value is, the better the quality of forecasts is.

3. Results and Discussion

In this Section we will present and compare the results obtained for the different approaches on the three considered stations. All the explicative variables are standardized in order to avoid the scale effect. For each station, we first give the variable ranking according to the two scores of importance *RFS* and ∂G_α, and then we select the subsets of relevant predictors using our stepwise algorithm. For each station the dataset is randomly splitted into training set (90 observations) and test set (13 observations). The training set is used for variable selection and small set is used for testing and selection bias checking.

It is well known that the selection bias problem is inherently related to variable selection tasks [33]. Indeed, when the test set is used to estimate the prediction error, then there will be a selection bias if this test set was used also in the variable selection process. The error rates obtained during the selection of the variables provides too-optimistic estimates. Thus, the test set must play no role in the variable selection process in order to obtain an unbiased error estimate.

3.1. *Experiments on the training sets*

At first, we have to tune the SVR parameters *d* (the degree of the polynomial kernel), ε and C. The grid search performed on several runs of 10-fold cross-validation gives rise to the results given in Table 3. We can see that all the datasets are nonlinear.

For the RF model, parameters *nodesize* and *mtry* are set to their default values for regression (*nodesize* = 5 and *mtry* = *p*/3, *p* is the number of variables) and we took *ntree* = 300 which lead to a good stability.

Table 3: SVR parameters tuning for the three stations

Station/Parameter	*D*		*C*
Gabes	2	0.01	1
Ghazela	2	0.001	1
Manouba	3	0.001	1

For each station, the training set contains 90 observations. The training sets are used to compute the variable importance according to the scores *RFS* and ∂G_α. Table 4 gives the Spearman's rank correlation coefficients ρ in order to measure the similarities between the different hierarchies across scores or/and stations.

Table 4: Spearman's rank correlation coefficients. Comparison of the hierarchies across scores or/and stations

		Gabes		Ghazela		Manouba	
		RFS	∂G_α	*RFS*	∂G_α	*RFS*	∂G_α
Gabes	*RFS*	1	(0.42)	0.29	0.30	0.12	0.22
	∂G_α		1	0.29	**0.54**	-0.04	**0.14**
Ghazela	*RFS*			1	(0.27)	0.40	0.06
	∂G_α				1	0.14	**0.03**
Manouba	*RFS*					1	(0.03)
	∂G_α						1

The underlined coefficients once measure the similarities between the *RFS* hierarchies across stations. The coefficients written in bold text measure the similarities between the ∂G_α hierarchies across stations. These similarity coefficients are qualified between medium and very low for both scores and more especially for the score ∂G_α. The coefficients in parentheses give the similarities across scores for each station. We see that Gabes station shows a relatively higher degree of similarity (ρ = 0.42) between the RFS and SVR rankings. Moreover, the obtained value of correlation coefficient is significant at the level of 99%. The remaining coefficients measure the similarities between the different hierarchies across scores and stations simultaneously. Finally, we can conclude that these hierarchies are very different but the *RFS* is a little bit more stable than the ∂G_α score from one station to the other.

Figures 2, 3 and 4 expose the variable ranking and the corresponding score value for Gabes, Ghazela and Manouba stations respectively.

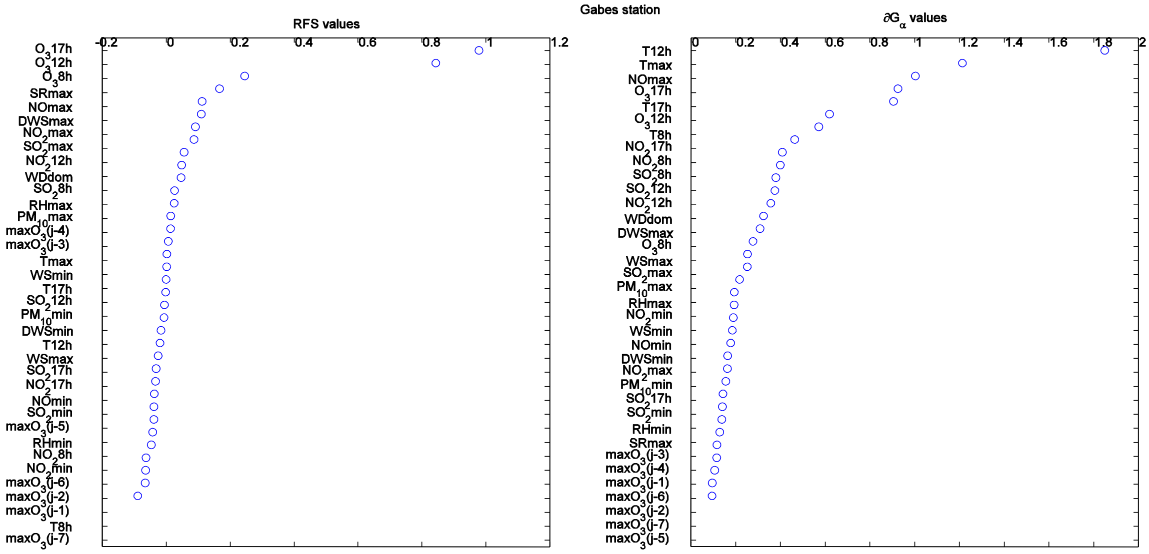

Figure 2: Variable ranking using *RFS* and ∂G_α scores for Gabes station

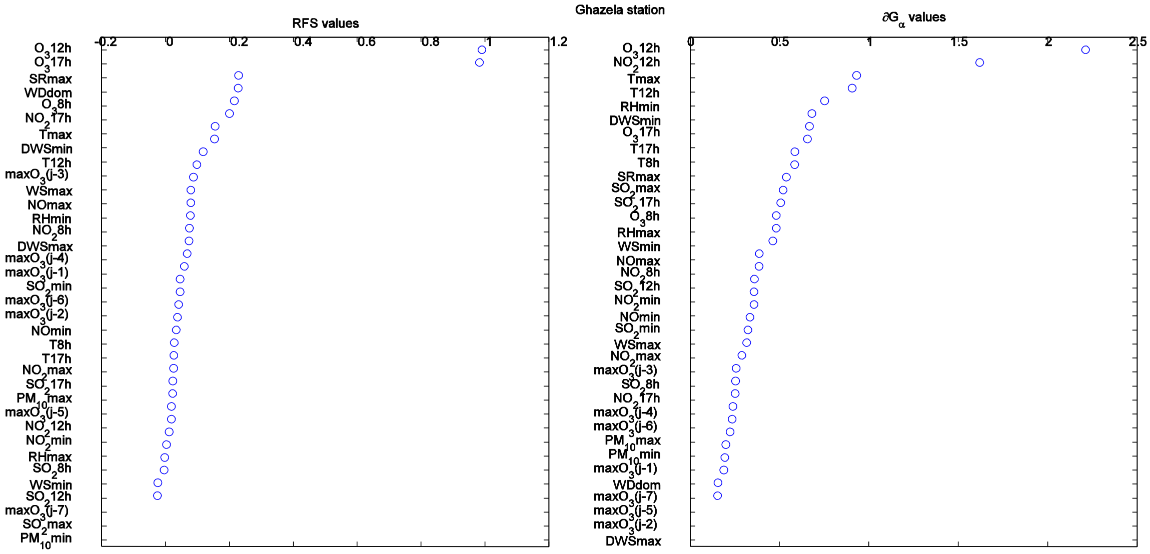

Figure 3: Variable ranking using *RFS* and ∂G_α scores for Ghazela station

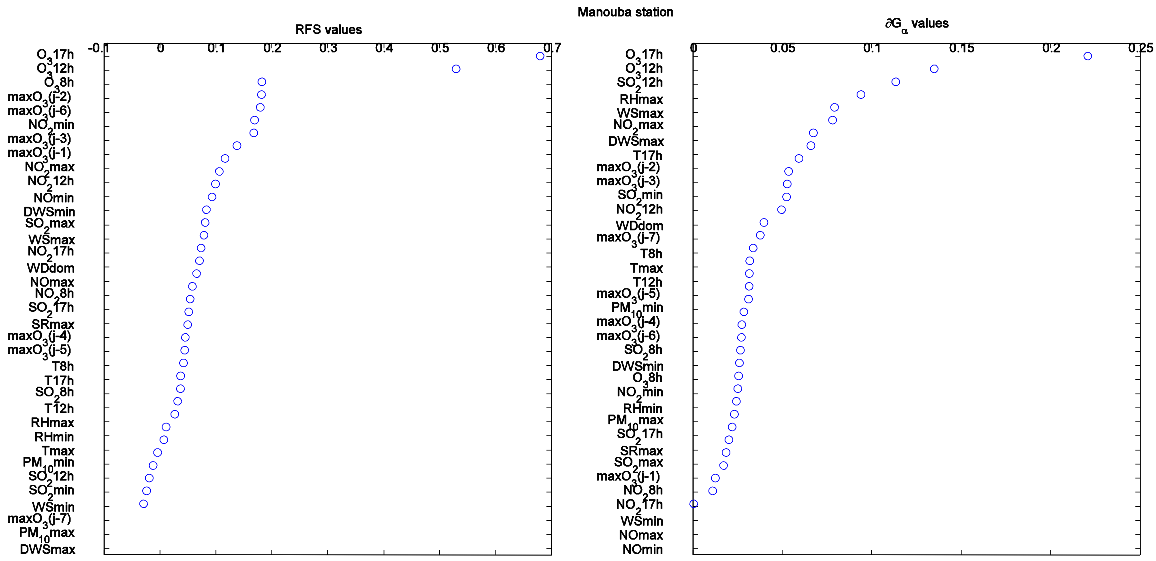

Figure 4: Variable ranking using *RFS* and ∂G_α scores for Manouba station

At first glance we can say that the *RFS* hierarchies are more admissible than those given by the ∂G_α score. Indeed, from these figures we note that the variables $O_3$8h, $O_3$12h and $O_3$17h occupy advanced ranks in the RFS hierarchies for the three stations. More specifically, the two variables $O_3$12h and $O_3$17h are strongly distinguished from the other. These findings are less true for the score ∂G_α. Moreover, even the hierarchies' headers produced by the score ∂G_α are extremely heterogeneous. Finally, we can see that the relevance of lagged maxO_3 is not high enough and differs greatly from one station to the other.

At this stage of investigation, we can say that the first step of variable ranking does not allow clear and fair comparison between the different approaches. Thus, the second step of selecting the optimal subset of variables will help us to complete our comparative study.

For the variable selection step, we will perform our stepwise algorithm with both RF and SVR using the two hierarchies given by the scores *RFS* and ∂G_α respectively. Using an external score to the model in the stepwise algorithm should reduce the selection bias problem [33]. The optimal subset of predictors is the one achieving the lowest *MSE* over 50 random splitting; 90% for learning and 10% for testing.

Figure 5 shows that, whatever the score, the stepwise algorithm performs well when using the RF model. Indeed, all the curves depicted in the top panels display the expected typical behavior; decreasing to reach a global minimum then increasing. This typical behavior reflects good performance of the stepwise algorithm and jointly attests good quality of the variable ranking [22,38]. This typical behavior is far from being realized with the SVR model especially when using the RFS hierarchy. The difficulty faced by the SVR model is more serious when dealing with Manouba dataset which is strongly nonlinear compared to those from Gabes and Ghazela stations. These results confirm that the *RFS* hierarchies are more plausible than those produced by the ∂G_α score for the three datasets.

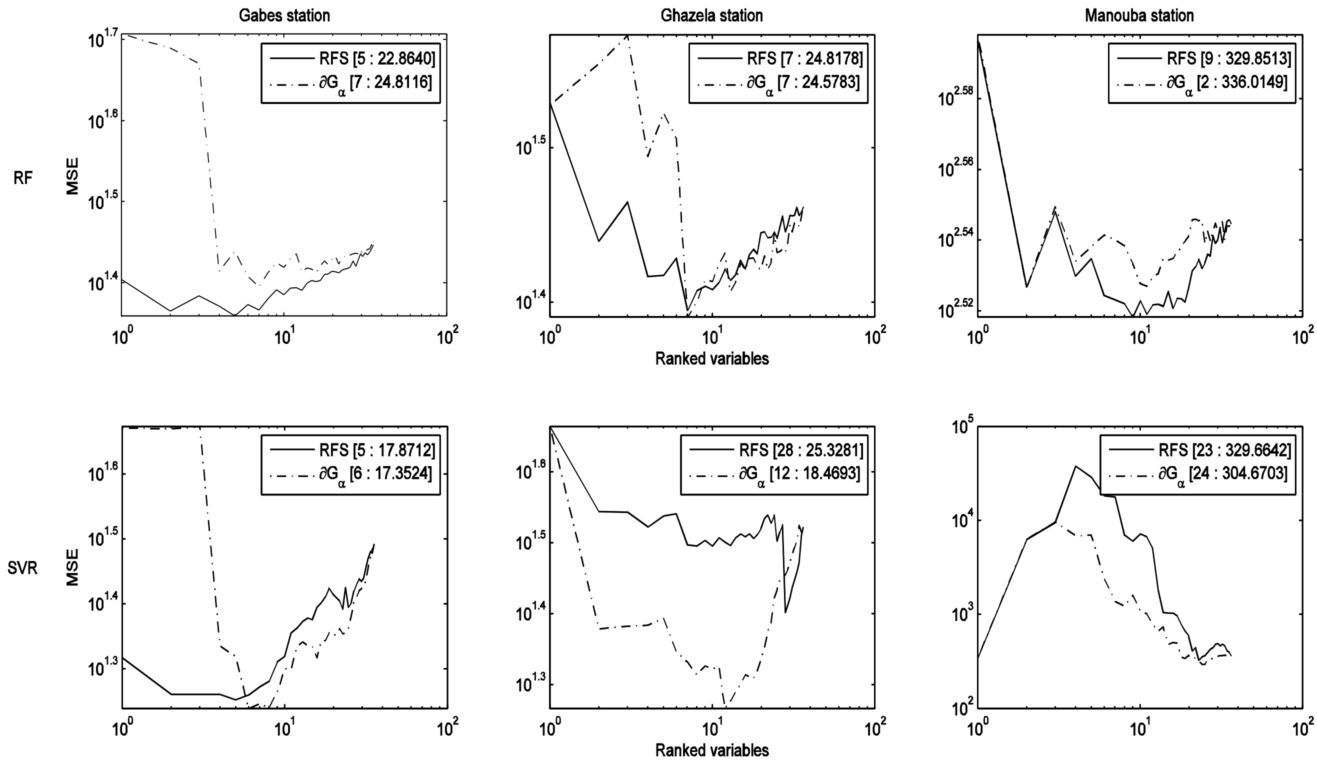

Figure 5: Mean Squared Error of nested increasing models. Each column of panels corresponds to a station and each row of panels corresponds to a model. For each curve, the optimal number of relevant predictors and the corresponding *MSE* are given in brackets. The *x* and *y* axes are taken in the logarithmic scale for clarity.

3.2. *Selection bias checking*

Finally, our goal now is to control the selection bias problem on the test sets, previously left aside. According to our previous analysis, we consider here only the hierarchies given by the *RFS*. The five, seven and nine top ranked variables are used for Gabes, Ghazela and Manouba datasets, respectively (see Figure 5). Only the RF model will be used for forecasting. To evaluate and to compare the forecasting effectiveness of the different models, we have adopted the statistical performance metrics given by the equations from (8) to (14). Table 5 gives the results achieved by the RF model when using all variables and when using only the selected ones. The best result for each criterion is written in bold text.

Table 5: RF-based *MSE* for the selected subsets of predictors and when using all the variables for the three stations

	RMSE	*MAE*	*MAPE*	FA_2	$FA_{1.25}$	d_r	E_1
				Gabes			
Selected	2.27	**1.82**	**3.94%**	**1**	**1**	**0.74**	**0.48**
All	**2.26**	1.85	4.08%	**1**	**1**	0.73	0.47
				Ghazela			
Selected	**4.02**	**2.97**	**3.22%**	**1**	**1**	**0.79**	**0.59**
All	4.16	3.18	3.51%	**1**	**1**	0.78	0.56
				Manouba			
Selected	**6.61**	5.48	8.70%	**1**	**1**	0.72	0.45
All	6.71	**5.29**	**8.63%**	**1**	0.92	**0.73**	**0.46**

From Table 5 we see that the forecasting accuracy is at least conserved by the variable selection for the three stations. The least significant improvement was achieved on Manouba station. However, almost all the used metrics were slightly improved by variable selection for Gabes and Ghazela stations. Moreover, These results demonstrate that our variable selection procedure is not affected by the selection bias problem and the RF model is robust against the overfitting problem. On the other hand, it is worthy to note that a decrease in at least one of the two measures *MAE* or *MAPE* is accompanied by an improvement in the last four metrics. This improvement becomes even more important when the decrease in *MAE* and/or *MAPE* is significant.

Ultimately, Figure 6 shows the forecasting performance of the selected subsets of predictors compared to using all the predictors. The observed values versus the forecasts of maxO_3 concentrations are depicted for each dataset. Each scatter plot corresponds to one station. We can see that the overall quality of forecasts is at least preserved when using only the selected predictors.

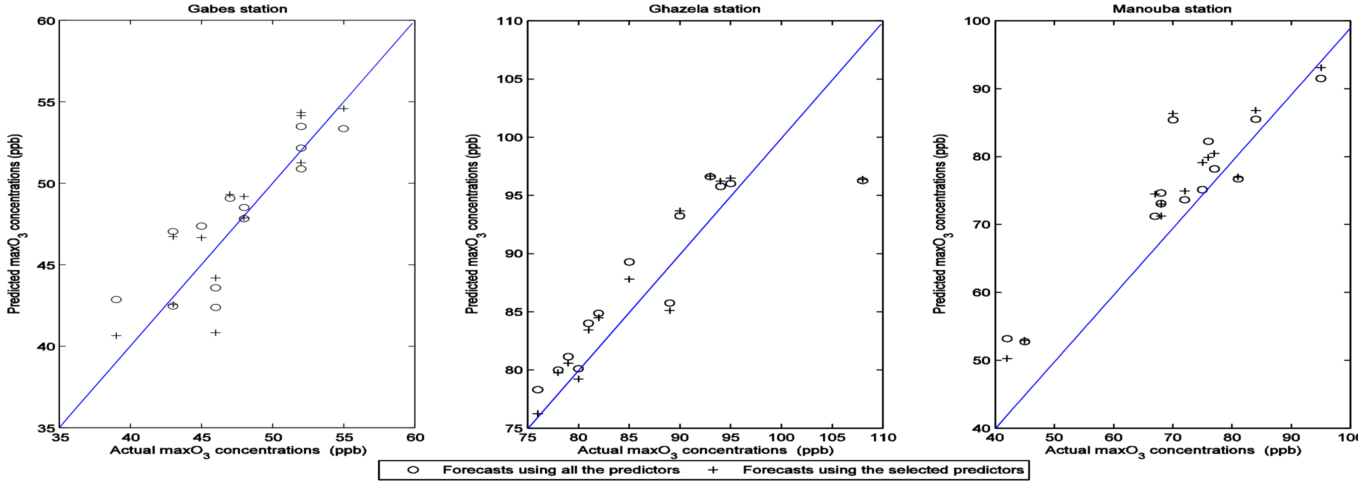

Figure 6: Actual values versus forecasts of maxO_3 concentrations using the RF model. Comparison between the forecasts using all the predictors and those using only the selected predictors for the three stations

Conclusions

We wanted this work to address a variety of researchers whatever their specialty, in hope that they found it useful. It is for this reason that we did not want to expose too technical and mathematical details and we rather preferred to focus on the applications.

In this work, we have compared two popular statistical learning models namely the Support Vector Regression and the Random Forests. We have considered three monitoring stations from Tunisia to model and to forecast the daily maximum ozone concentration maxO_3. These three stations reflect the diversity of urban situations; background, traffic and industrial cities. The problem of variable selection for regression was deeply investigated.

Methodologically, we have shown that RF outperforms SVR in variable importance assessment and in variable selection. The SVR model has encountered more difficulties on Manouba station dataset which was heavily nonlinear. One of the major drawbacks of SVR is the limited choice of kernel functions and their parameters tuning. However, RF are highly nonparametric statistical tool which handle data without need to transform them beforehand. Thereby, in the stepwise procedure the RF model fits automatically to the data at each variable introduction. Unfortunately, the SVR model does not have this flexibility quality. Moreover, we have shown that the RF model is not affected by the problem of selection bias.

In practice, we have shown that it is possible to accurately forecast the daily maximum ozone concentration maxO_3 by using a small subset of selected variables. We have found that the ozone concentrations measured at particular periods of the day are crucial to accurately forecast the current day maxO_3 value. This result is confirmed on the three stations despite their urban, meteorological and geographic great differences. Finally, this work could be broadened to study other pollutants from various monitoring stations in different cities.

Acknowledgments- The authors would like to acknowledge the Tunisian National Agency for Environmental Protection (ANPE) for providing the data.

References

1. Cakmak S., Hebbern C., Cakmak J.D., Vanos J., The modifying effect of socioeconomic status on the relationship between traffic, air pollution and respiratory health in elementary schoolchildren, *J. Environ. Manage.* 177 (2016) 1–8.
2. Agirre-Basurko E., Ibarra-Berastegui G., Madariaga I., Regression and multilayer perceptron-based methods models to forecast hourly O3 and NO2 levels in the Bilbao area, *Environ. Model. Softw.* 21 (2006) 430–446.
3. Rahman S.M., Khondaker A.N., Abdel-Aal R., Self organizing ozone model for Empty Quarter of Saudi Arabia: Group method data handling based modeling approach, *Atmos. Environ.* 59 (2012) 398–407.
4. Guerra J.C., Rodríguez S., Arencibia M.T., García M.D., Study on the formation and transport of ozone in relation to the air qualitymanagement and vegetation protection in Tenerife Canary Islands, *Chemosphere* 56 (2004) 1157–1167.
5. Zolghadri A., Monsion M., Henry D., Marchionini C., Petrique O., Development of an operational model-based warning system for tropospheric ozone concentrations in Bordeaux, France, *Environ. Model. Softw.* 19 (4) (2004) 369–382.
6. Bytnerowicz A., Omasa K., Paoletti E., Integrated effects of air pollution and climate change on forests: a northern hemisphere perspective, *Environ. Pollut.* 147 (3) (2006) 438–445.
7. Pires J.C.M., Sousa S.I.V., Pereira M.C., Alvim-Ferraz M.C.M., Martins F.G., Management of air quality monitoring using principal component and cluster analysis e part II: CO, NO2 and O3, *Atmos. Environ.* 42 (6) (2008) 1261–1274.
8. Ortiz-García E.G ., Salcedo-Sanz S., Pérez-Bellido Á.M., Portilla-Figueras J.A., Prieto L., Prediction of hourly O_3 concentrations using support vector regression algorithms, *Atmos. Environ.* 44 (35) (2010) 4481–4488.
9. Windsor H.L., Toumi R., Scaling and persistence of UK pollution, *Atmos. Environ.* 35 (2001) 4545–4556.
10. Dueñas C., Fernandez M.C., Canete S., Carretero J., Liger E., Assessment of ozone variations and meteorological effects in an urban area in the Mediterranean Coast, *Sci. Total Environ.* 299 (2002) 97–113.
11. Pires J.C.M., Alvim-Ferraz M.C.M., Pereira M.C., Martins F.G., Prediction of tropospheric ozone concentrations: application of a methodology based on the Darwin's theory of evolution, *Expert Syst Appl* 38 (2011) 1903–1908.
12. Borrego C., Tchepel O., Costa A.M., Amorim J.H., Miranda A.I., Emission and dispersion modeling of Lisbon air quality at local scale, *Atmos. Environ.* 37 (2003) 5197–5205.
13. Yi J., Prybutok V.R., A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area, *Environ. Pollut.* 92 (3) (1996) 349–357.
14. Spellman G., An application of artificial neural networks to the prediction of surface ozone concentrations in the United Kingdom, *Appl. Geogr.* 19 (2) (1999) 123–136.
15. Gómez-Sanchis J., Martín-Guerrero J.D., Soria-Olivas E., Vila-Francés J., Carrasco J.L., del Valle-Tascón S., Neural networks for analysing the relevance of input variables in the prediction of tropospheric ozone concentration, *Atmos. Environ.* 40 (32) (2006) 6173–6180.
16. Feng Y., Zhang W., Sun D., Zhang L., Ozone concentration forecast method based on genetic algorithm optimized back propagation neural networks and support vector machine data classification, *Atmos. Environ.* 45 (2011) 1979–1985.
17. Wang P., Liu Y., Qin Z., Zhang G., A novel hybrid forecasting model for PM_{10} and SO_2 daily concentrations, *Sci. Total Environ.* 505 (2015) 1202–1212.

18. Genuer R., Poggi J.M., Tuleau C., Variable selection using random forests, *Pattern Recognit Lett* 31 (14) (2010) 2225–2236.
19. Yang Z.C., Modeling and forecasting daily movement of ambient air mean $PM_{2.5}$ concentration based on the elliptic orbit model with weekly quasi-periodic extension: a case study, *Environ. Sci. Pollut. R.* 21 (16) (2014) 9959–9972.
20. Antanasijević D.Z., Pocajt V.V., Povrenović D.S., Ristić M.Đ., Perić-Grujić A.A, PM_{10} emission forecasting using artificial neural networks and genetic algorithm input variable optimization, *Sci. Total Environ.* 443 (2013) 511–519.
21. Al-Alawi S.M., Abdul-Wahab S.A., Bakheit C.S., Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone, *Environ. Model. Softw.* 23 (2008) 396–403.
22. Ghattas B., Ben Ishak A., Sélection de variables pour la classification binaire en grande dimension: comparaisons et application aux données de biopuces, *J-SFdS* 149 (3) (2008) 43–66.
23. Ghattas B. Prévision des pics d'ozone par arbres simples et agrégés par bootstrap, *Rev Stat Appl* Vol. XLVII (2) (1999) 61–80.
24. Pavón-Domínguez P., Jiménez-Hornero F.J., Gutiérrez-de-Ravé E., Proposal for estimating ground-level ozone concentrations at urban areas based on multivariate statistical methods, *Atmos. Environ.* 90 (2014) 59–70.
25. Debry É., Mallet V., Ensemble forecasting with machine learning algorithms for ozone, nitrogen dioxide and PM10 on the Prev'Air platform, *Atmos. Environ.* 91 (2014) 71–84.
26. Van Buuren S., Groothuis-Oudshoorn K., mice: Multivariate Imputation by Chained Equations in R, *J Stat Softw* 45 (3) (2011).
27. Vapnik V.N., Golowich S., Smola A., Support vector method for function approximation regression estimation and signal processing, *In M. Mozer M. Jordan and T. Petsche editors, Advances in Neural Information Processing Systems pages*, Cambridge, MA, MIT Press, (1997).
28. Vapnik V.N., *Statistical learning theory*, Wiley, New York, (1998).
29. Breiman L., Random Forests, *Mach Learn* 45 (1) (2001) 5–32.
30. Ben Ishak A., Variable selection using support vector regression and random forests: A comparative study, *Intell Data Anal* 20 (1) (2016) 83–104.
31. Rakotomamonjy A., Analysis of SVM regression bounds for variable ranking, *Neurocomputing* 70 (7–9) (2007) 1489–1501.
32. Chang M.W., Lin C.J., Leave-one-out bounds for support vector regression model selection, *Neural Comput* 17 (4) (2005) 1–26.
33. Ambroise C., McLachlan G.J., Selection bias in gene extraction on the basis of microarray gene-expression data, *PNAS* 99 (10) (2002) 6562–6566.
34. Legates D.R., McCabe G.J., A refined index of model performance: a rejoinder, *Int. J. Climatol.* 33 (4) (2013) 1053–1056.
35. Willmott C.J., Robeson S.M., Matsuura K., A refined index of model performance, *Int. J. Climatol.* 32 (13) (2012) 2088–2094.
36. Krause P., Boyle D.P., Bäse F., Comparison of different efficiency criteria for hydrological model assessment, *Adv. Geosci.* 5 (2005) 89–97.
37. Koo Y.-S., Kim S.-T., Cho J.-S., Jang Y.-K., Performance evaluation of the updated air quality forecasting system for Seoul predicting PM_{10}, *Atmos. Environ.* 58 (2012) 56–69.
38. Feki A., Ben Ishak A., Feki S., Feature selection using Bayesian and multiclass support vector machines approaches: Application to bank risk prediction, *Expert Syst Appl* 39 (3) (2012) 3087–3099.

(2017) ; http://www.jmaterenvironsci.com/