ISEIS

Journal of Environmental Informatics 30(2) 79-94 (2017)

Journal of Environmental Informatics

www.iseis.org/jei

Variable Selection Based on Statistical Learning Approaches to Improve PM_{10} Concentration Forecasting

A. Ben Ishak*

Université de Tunis, ISGT, LR99ES04 BESTMOD, Le Bardo 2000, Tunisia

Received 02 Feb 2015; revised 07 Jul 2015; accepted 27 Jul 2015; published online 19 Sep 2016

ABSTRACT. In this work, the problem of variable selection for regression is investigated in order to improve the forecasting accuracy. To that effect, the support vector regression (SVR) and the random forests (RF) are used to assess the variable importance. Then, a stepwise algorithm is built to select the best subset of predictors. An intensive comparative study is conducted on simulated and real datasets. The real datasets expose the problem of particulate matter concentration forecasting in two monitoring stations from Tunisia. We have proposed a combined approach using SVR and RF for variable importance assessment and for variable selection. We have achieved a significant improvement in forecasts accuracy for the two stations when using only a reduced number of selected predictors.

Keywords: support vector regression, random forests, variable selection, stepwise algorithm, selection bias, particulate matter forecasting

1. Introduction

Air pollution is a fundamental problem in many parts of the world. It is usually caused by energy production from power plants, industrial processes, residential heating, fuel burning vehicles, natural disasters, etc. Consequently, there is a growing interest, in day-to-day, in air quality surveillance. 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 (Ortiz-García et al., 2010). 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. General Overview on PM_{10}

Specifically, particulate matter (PM) is a widespread air pollutant, consisting of a mixture of solid and liquid particles suspended in the air. The mass concentration of particles with a diameter of less than 10 μm is commonly noticed by PM_{10} and of particles with a diameter of less than 2.5 μm is commonly called $PM_{2.5}$. The particles $PM_{2.5}$, often called fine particulate matter, also comprises ultrafine particles having a diameter of less than 0.1 μm.

Particles can either be directly emitted into the air (primary PM) or be formed in the atmosphere from gaseous precursors such as sulfur dioxide, oxides of nitrogen, ammonia and non-methane volatile organic compounds (secondary particles). Primary PM and the precursor gases can have both man-made (anthropogenic) and natural (non-anthropogenic) sources. Secondary particles are formed in the air through chemical reactions of gaseous pollutants. They are products of atmospheric transformation of nitrogen oxides (mainly emitted by traffic and some industrial processes) and sulfur dioxide resulting from the combustion of sulfur-containing fuels. Secondary particles are mostly found in fine PM.

Particulate matter PM_{10} is one of the pollutant that have harmful effects on human health and environment. PM_{10} is commonly considered as one of the major factors that contributes to air pollution problems (Pope III, 2000; Moshammer and Neuberger, 2003; Hauck et al., 2004). It is well established nowadays that short and long term exposure to high particulate matter concentrations causes adverse health effects and reduction in population's life expectancy in both developed and developing countries. A number of toxicological and epidemiological studies reported that exposure to particulate matter can cause health problems ranging from respiratory to cardiovascular illnesses (Pope III and Dockery, 2006; Perez et al., 2009; Russell and Brunekreef, 2009).

1.2. Related Works

To forecast pollution concentrations often are used: statistical models, methods based on reasoning rules, neural networks, filtering of time series, support vector machines, clustering ana-

* Corresponding author. Tel.: +216 97 549940; fax: +216 71 588487.
E-mail address: anis_isg@yahoo.fr (A. Ben Ishak).

ISSN: 1726-2135 print/1684-8799 online
 doi:10.3808/jei.201600341

lysis, etc. These methods have the possibility of discovering new dependencies between data gathered in sets. However, in the recent years, pollution concentrations forecasts are more often based on the data mining methods.

In environmental sciences, a lot of research efforts go towards the understanding of air quality phenomenon and the ability to forecast it. In the recent years, various parametric and nonparametric statistical models have been developed to analyze and predict PM_{10} emission and dispersion. First of all, neural networks are the most frequently used to produce forecasts of PM_{10} concentration (Kukkonen et al., 2003; Paschalidou et al., 2009; Carnevale et al., 2011). Therefore, multiple linear regression modeling was also employed as in the studies of Stadlober et al. (2008), Cordelino et al. (2001) and Paschalidou et al. (2009). Moreover, Chaloulakou et al. (2003) and Grivas and Chaloulakou (2006) compared the performance of neural networks and multiple regression model to forecast the daily average of PM_{10} concentration. Corani (2005) used local polynomial based nonparametric approach to estimate a nonlinear regression model in Milan where the PM_{10} pollution is important. Hoi et al. (2009) proposed a time varying autoregressive model with exogenous input based on a Kalman filter in order to predict daily PM_{10} concentration. In the work of Slini et al. (2006) three approaches were compared for PM_{10} forecasting namely, the linear regression models, the Classification and Regression Trees (CART) and the artificial neural networks. More recently, PM_{10} concentration was predicted using cluster-wise linear models in the studies of Sfetsos and Vlachogiannis (2010), and Poggi and Portier (2011). A bit further, a combination of Mesoscale Model (MM5) and Community Multi-scale 3-D Air Quality modeling system was employed to investigate the PM_{10} pollution issue in Beijing, China (Huang et al., 2010; Zhou et al., 2012). The authors have focused on the effects of different restriction policies implemented during and after the 2008 Olympic Games, and investigated the PM_{10} source apportionment.

Although a variety of statistical tools have been used in previous studies to forecast the PM_{10} concentration, few researchers have focused on the explanatory variables significance. However, the PM_{10} concentration is influenced by many meteorological factors and primary pollutants, while the influence ability and the influence degree are uncertain. More broadly, understanding, modeling and forecasting the PM_{10} concentration is difficult due to the distinctive influence of the area topography, geomorphology, emission source location, discharged rate and meteorological factors. Moreover, the forecasting difficulty is principally due to the complex characteristics of data. The works of Antanasijević et al. (2013) and Qin et al. (2014) are among the few studies that address the variable importance issue. In the first work, the authors constructed a nonlinear predictive model based on gray correlation analysis (GCA), Ensemble Empirical Mode Decomposition (EEMD), Cuckoo search (CS) and Back-propagation artificial neutral networks (BPANN) to identify the pertinent predictors giving rise to an accurate model to forecast the PM_{10} concentrations in different climatic and environmental areas. Variable importance evaluation was done using some kind of linear correlation coefficient. Their analysis results indicate that air pollutants are more closely related to PM_{10} than to meteorological factors. Yang (2014) disagreed with these results and argued that emission sources and meteorological conditions can strongly govern the spatial and temporal variability of air pollutant concentration and its daily movement. The second work is quite different from the first one. Indeed, Antanasijević et al. (2013) developed an artificial neural network (ANN) model for the forecasting of annual PM_{10} emissions at the national level, using economical/industrial parameters as explicative variables. It was shown that the selection of inputs, based on smoothing factor calculated by genetic algorithm, provides much accurate forecasts in comparison with conventional models.

More recently, Wang et al. (2015) proposed a hybrid methodology based on ANN and support vector machines (SVM) coupled with the Taylor expansion forecasting model. The proposed hybrid method shows superior accuracy in PM_{10} and SO_2 forecasting results. Without performing variable selection in their work, they stated that more complete input variables are required to obtain more accurate results, and the selection of the input variables influences the accuracy of forecasting.

1.3. Objectives and Outline of This Work

This research is mainly motivated by the desire to fill the existing literature lack on variable selection issue for PM_{10} concentration forecasting. Indeed, we used machine learning approaches to assess variable importance and to overcome the complex characteristics of data. To this end, we compare two popular statistical learning approaches: the support vector regression and the random forests (respectively, SVR and RF henceforth). The random forests are nowadays one of the most powerful learning methods, and the support vector regression approach has been successfully used in a wide variety of applications. Also, we use a sequential strategy to select the most important variables firstly on toy data, and then to identify the most explicative predictors on two real datasets involving daily PM_{10} concentration.

In this study, we consider PM_{10} daily average concentrations measured in Gabes and Manouba, located in Tunisia. The Tunisian authorities monitor air pollution by means of the National Network for Monitoring Air Quality (RNSQA, French acronym of national network for air quality monitoring). This network contains 15 fixed monitoring stations. 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. Otherwise, in the northeastern Tunisia, Manouba is a polluted urban region characterized by the presence of some industries and important vehicular traffic.

The main contributions of this paper are as follows. First, we compare recent statistical learning methods for variable importance assessment and for variable selection. Second, we propose a mixed cooperative procedure between SVR and RF to determine the most explicative variables for the PM_{10} daily ave-

rage concentration in order to attenuate the selection bias. Moreover, this study is presumably the pioneering work that attempts to compare the SVR and the RF in variable selection for regression and uses them together to forecast the PM_{10} pollutant.

The remainder of this paper is organized as follows. First, Section 2 presents the used models, the variable importance assessment and the variable selection strategy. Furthermore, the real datasets are described in Section 3. The experimental results on simulated and real datasets are provided in Section 4. Finally, Section 5 is devoted to some concluding remarks and presents some possible perspectives.

2. Used Methods

In this section we briefly describe the main characteristics of the statistical tools used in this work without going into too much technical details. We focused on two popular and competitive approaches from the statistical learning literature, namely support vector regression and random forests. Then we shortly depict some SVR scores and the RF score for the purpose of variable importance assessment. Finally, we present our stepwise algorithm.

2.1. Variable Importance Using Support Vector Machines

Support vector machines approach is a relatively popular computational learning algorithm based on the statistical learning theory developed by Vapnik (1995, 1998). The foundations of support vector machines gained popularity due to many promising features such as better empirical performance even on a limited number of learning patterns. Support vector machines were originally designed to solve biclass problems (Boser et al., 1992; Cristianini and Taylor, 2000). The ingenious idea of support vector machines has been recently extended by Vapnik et al. (1997) to handle regression problems.

2.1.1. Support Vector Machines for Regression

Suppose we have a given learning dataset $\{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)\} \subset X \times Y$ where $X \subset R^p$ is the input space, p is the number of explanatory variables and $Y \subset R$ is the output space. In the support vector regression formulation, the goal is to find a linear function $f(x) = \langle \boldsymbol{w} \mid \varphi(\boldsymbol{x}) \rangle_H + b$ in a reproducing kernel Hilbert space H, commonly called feature space. The weight vector $\boldsymbol{w}$ and the pattern $\varphi(\boldsymbol{x})$ belong to H, the real number b is called the bias, the operator $\langle \cdot \mid \cdot \rangle$ denotes the standard inner product and the function φ: $X \rightarrow H$ is an implicit nonlinear mapping induced by a kernel function K satisfying Mercer's conditions. When the targets y_i and the patterns x_i are not linearly correlated in the input space, the kernel function K is used to embed the data into a higher dimensional feature space where in general it is guaranteed that a linear regression function may be found (Vapnik, 1995, 1998). That function should have at most ε-deviation from the real targets y_i for all the mapped training vectors $\varphi(\boldsymbol{x}_i)$ and at the same time is as flat as possible.

The L2-SVR algorithm looks for minimizing the following regularized risk (Vapnik, 1998; Smola and Schölkopf, 1998; Shawe-Taylor and Cristianini, 2004):

$$R[f] = \frac{1}{2} \| \boldsymbol{w} \|^2 + C \sum_{i=1}^{n} L(x_i, y_i, f) \tag{1a}$$

where the hyperparameter C determines the tradeoff between the flatness of f and the amount up to which deviations larger than ε are tolerated, and $L(x_i, y_i, f)$ is the quadratic ε-insensitive loss function described by:

$$L(x_i, y_i, f) = \max\{0, (| y_i - f(\boldsymbol{x}_i) | - \varepsilon)^2\} \tag{1b}$$

It turns out that the convex optimization problem can be solved more easily in its dual formulation:

$$\underset{\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}}{maximize} \sum_{i=1}^{n} y_i (\hat{\alpha}_i - \alpha_i) - \varepsilon \sum_{i=1}^{n} (\hat{\alpha}_i + \alpha_i) - \frac{1}{2} \sum_{i,j=1}^{n} (\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j) (K(\boldsymbol{x}_i, \boldsymbol{x}_j) + \frac{1}{c} \delta_{ij}) \tag{1c}$$

subject to

$$\sum_{i=1}^{n} (\hat{\alpha}_i - \alpha_i) = 0 \tag{1d}$$

$$\hat{\alpha}_i \geq 0, \alpha_i \geq 0, \forall i = 1, 2, ..., n \tag{1e}$$

where $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \varphi(\boldsymbol{x}_i) \mid \varphi(\boldsymbol{x}_j) \rangle$ is the used kernel, α_i and $\hat{\alpha}_i$ for $i = 1, 2, \ldots, n$ are the Lagrangian multipliers, and δ_{ij} is the Kronecker symbol.

The solution of this quadratic convex problem can be obtained by means of Lagrangian theory and it gives rise to the following expansion:

$$\boldsymbol{w} = \sum_{i,j=1}^{n} (\hat{\alpha}_i - \alpha_i) \varphi(\boldsymbol{x}_i) \tag{1f}$$

Hence, we arrive at the SVR function given by:

$$f(x) = \sum_{i,j=1}^{n} (\hat{\alpha}_i - \alpha_i) K(\boldsymbol{x}_i, \boldsymbol{x}) + b \tag{1g}$$

The two widely used families of kernels are the polynomial and the Gaussian.

2.1.2. Variable Importance in SVR

The statistical theoretical wealth and the structure of support vector machines for classification allow to estimate their generalization performance from bounds on the leave-one-out error, which is known to be an almost unbiased estimator of the

expected generalization error. The two most frequently used bounds are the so-called radius-margin bound established by Vapnik (1998) and the span-bound given by Vapnik & Chapelle (2000).

For the regression framework, Chang and Lin (2005) have extended these results for SVR. More recently, Rakotomamonjy (2007) derived from these bounds some criteria for variable importance assessment purpose.

For the sake of being self-contained and concise at the same time, we only summarize here the different criteria that will be used for the sequel of the paper. For more mathematical details about the scores computations by means of derivatives, we suggest the reader to refer to the works of Chang and Lin (2005) and Rakotomamonjy (2007).

In our application part, the following four criteria will be used. The two first criteria are bounds on the leave-one-out error so they are directly related to the predictive performance of the SVR:

- Radius-margin bound: $G_R(\boldsymbol{\alpha},\hat{\boldsymbol{\alpha}}) = R^2 \sum_{i=1}^{n} (\hat{\alpha}_i + \alpha_i)$ where R^2 is the radius of the smallest sphere containing the set of mapped patterns $\{\varphi(\boldsymbol{x}_i)\}_{i=1, 2, \ldots, n}$.
- Span-bound: $G_S(\boldsymbol{\alpha},\hat{\boldsymbol{\alpha}}) = \sum_{i=1}^{n} (\hat{\alpha}_i + \alpha_i) \tilde{S}_i^2$, where $\tilde{S}_i^2$ is the regularized version of the squared distance of $\varphi(\boldsymbol{x}_i)$ to the span of all other support vectors.

The coming two criteria are not bounds, but they were proposed by Rakotomamonjy (2007) as a supplementary criteria for variable ranking because they are relatively cheaper to compute:

- $G_\alpha(\boldsymbol{\alpha},\hat{\boldsymbol{\alpha}}) = \sum_{i=1}^{n} (\hat{\alpha}_i + \alpha_i)$: is a principal term involved in both radius-margin and span estimate bounds.
- $G_W(\boldsymbol{\alpha},\hat{\boldsymbol{\alpha}}) = \sum_{i,j=1}^{n} (\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j)(K(\boldsymbol{x}_i,\boldsymbol{x}_j) + \delta_{ij} / C)$: is the norm of the regression function f in the feature space H.

The four scores derived from these criteria will be denoted in this work by ∂G_R, ∂G_S, ∂G_α and ∂G_W, respectively. Each score may be computed from an SVR model learned from the available dataset. To get better estimates of the scores, we computed for them a bootstrap estimate in a way similar to that in the bagging method (Breiman, 1996). Bootstrapping the scores seems to be more robust to data variation and attenuates the effect of selection bias (Ambroise and McLachlan, 2002; Ben Ishak and Ghattas, 2005). Once the scores are computed, all the variables may be ranked in a decreasing order of importance.

2.2. Random Forests

Random Forests, introduced by (Breiman, 2001), are nowadays one of the most popular and successful learning methods in both classification and regression. They received much attention due to their computational fastness and remarkable empirical success. In this work, we will focus on random forests for regression.

2.2.1. Model Presentation

A forest is an ensemble of trees like in real life. Breiman (2001) introduced the general concept of random forests based on binary decision trees (Classification And Regression Trees, CART, (Breiman et al., 1984)).

Rather than using a single tree, random forests construct an ensemble predictor by averaging over a collection of binary trees. A forest is random in two ways: (i) each tree is grown from an independent bootstrap sample of the data, and (ii) at each node of the tree a randomly selected variables are chosen as candidate variables to split on. The two main parameters of random forests are *mtry*, the number of input variables randomly chosen at each split and *ntree*, the number of trees in the forest. A third parameter, denoted by *nodesize*, is the minimal size of the leaves of the trees. We retain the default value (5 for regression) in our experimentations, since it is close to the maximal tree choice.

Trees are quite unstable, so that this randomness creates differences in individual trees' predictions. This enables random forests to adapt to the data, automatically fitting higher order interactions and nonlinear effects, while at the same time keeping overfitting in check. This has led to a great interest in the method and its application to many fields.

2.2.2. Variable Importance in RF

While random forests are often used for exploratory data analysis, they can also be used to select variables and reduce dimensionality. This is done by ranking variables by some measure of individual importance. More recently, Gregorutti et al. (2015) adapted the individual importance measure for groups of variables.

In the random forests framework for regression problems, the most widely used score of importance of a given variable, suggested by (Breiman, 2001), 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*"). According to random sampling of observations, there are on average $1/e \approx 36.8\%$ of the observations that are not used for building the current tree; that is, are Out-Of-Bag for that tree. The accuracy of a random forest's prediction can be estimated from these *OOB* data as:

$$OOB_{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \overline{\hat{y}}_{i_{OOB}} \right)^2 \qquad (2)$$

where $\overline{\hat{y}}_{i_{OOB}}$ denotes the average prediction for the ith observation from all trees for which this observation has been *OOB*. To avoid insignificant sampling effects, each *OOB* error is the mean of *OOB* errors over several runs.

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

- For each tree $t = 1, 2, \ldots,$ *ntree* in the forest the *OOB* Mean Squared Error is computed. It is the average of the squared deviations of *OOB* responses from their respective predictions:

$$OOB_{MSE}^{t} = \frac{1}{\left|OOB^{t}\right|} \sum_{i \in OOB^{t}} \left(y_i - \hat{y}_{i,t}\right)^2 \quad (3a)$$

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

- For each variable $j = 1, 2, \ldots, p$, the *OOB* Mean Squared Error is computed for each tree $t = 1, 2, \ldots, ntree$ on the associated perturbed *OOB* sample, OOB^t, by randomly permuting the values of the j^{th} variable:

$$\tilde{OOB}_{MSE}^{t\,j} = \frac{1}{\left|\tilde{OOB}^{t\,j}\right|} \sum_{i \in \tilde{OOB}^{t\,j}} \left(y_i - \hat{y}_{i,t}\right)^2 \quad (3b)$$

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

$$\tilde{OOB}_{MSE}^{t\,j} - OOB_{MSE}^{t} \quad (3c)$$

This difference is null for a variable that happens to be not involved in any split of tree t.

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

$$RFS_j = \frac{1}{ntree} \sum_{i=1}^{ntree} \left(\tilde{OOB}_{MSE}^{t\,j} - OOB_{MSE}^{t}\right) \quad (3d)$$

Unlike SVM, the RF importance score does not need bootstrapping because it is stable in the presence of noise and correlated variables and vis-a-vis to small perturbations of the data. But still to avoid insignificant sampling effects, RF importance score is the average on several runs.

Once all the variables are ranked in a decreasing order of importance, we apply a stepwise forward strategy in order to select the subset of the most explicative variables.

2.3. Variable Selection Algorithms

In supervised learning problems, variable selection is an important step for the training phase (Guyon et al., 2002; Rakotomamonjy, 2003; Guyon and Elisseeff, 2003). The goal is to select an "optimal" subset of variables in order to improve or, at least, to preserve the predictive performance of the model. Thus, the motivation for feature selection is improving prediction accuracy, reducing time complexity and facilitating data understanding.

To this end, one needs an efficient algorithm for selecting an "optimal" subset of variables. This search-space optimization problem is NP-hard (Amaldi and Kann, 1998), so one has to rely on approximation strategies.

Feature selection methods may differ according to the nature of the evaluation criteria; wrapper, filter or embedded. The wrapper method, integrates the model prediction performance into the evaluation of the quality of a subset of features. The filter method evaluates the variable importance by using a statistic criterion independent of the model. The embedded method combines feature selection and model prediction into one task.

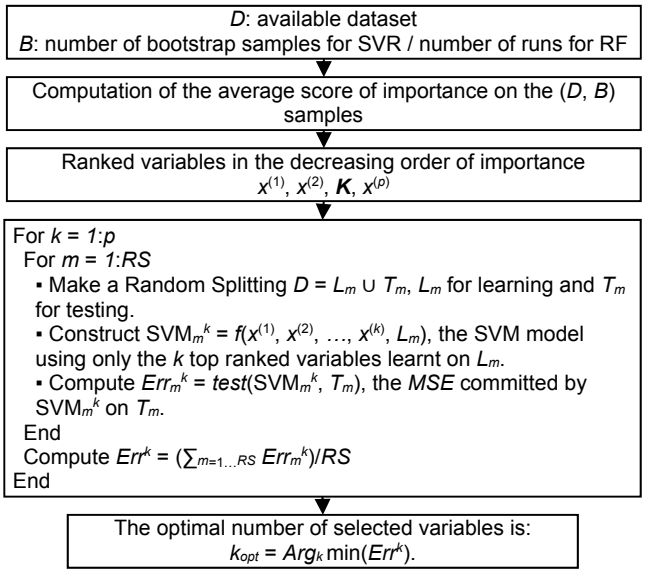

Figure 1. The variable selection procedures.

In this work, we consider a feature selection algorithm from the wrapper category using SVR and RF models. We choose a stepwise forward strategy, firstly introduced for classification in the work of Ghattas and Ben Ishak (2008), based on a sequential introduction of variables. A sequence of nested increasing models $M^{(k)}$, $k = 1, 2, \ldots, p$, 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 random splitting for the SVM and estimated both by random splitting and on *OOB* samples for the RF. The set of variables leading to the model of smallest error rate is selected. Our stepwise algorithm has shown a promising results on classification problems even in the situations exposing the curse of dimensionality phenomenon, i.e., when the number of input variables p is very large compared to the sample size n (Ghattas and Ben Ishak, 2008; Feki et al., 2012).

The detailed main steps of our stepwise procedure, using the SVM models with random splitting, are depicted in the flowchart given by Figure 1. The term SVM_m^k is replaced by RF_m^k when dealing with the RF models using random splitting. The internal *For* loop is removed when the error rate is estimated on *OOB* samples because it is a built-in default task in the RF model.

3. Real Data Gathering and Pretreatments

Here, we expose the real data collection, the explicative variable description and the missing values treatment.

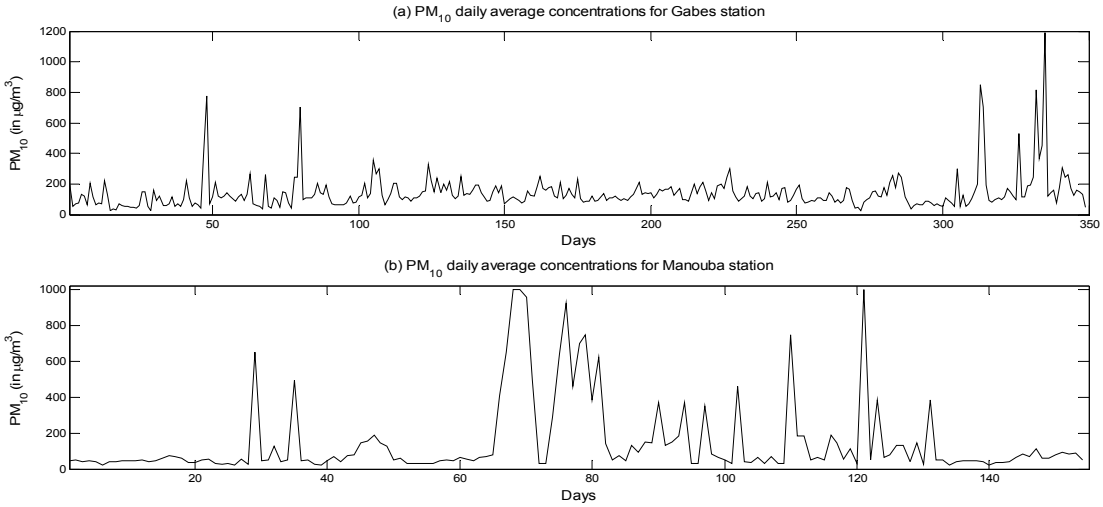

Figure 2. Variation of PM_{10} daily average concentrations: (a) Gabes station; (b) Manouba station.

Table 1. Twenty-four Explicative Variables

Type	Variable	Definition
Meteorological predictors	Tmin	Daily minimum temperature (°C)
	Tmoy	Daily mean temperature (°C)
	Tmax	Daily maximum temperature (°C)
	VVmin	Daily minimum wind speed (m/s)
	VVmoy	Daily mean wind speed (m/s)
	VVmax	Daily maximum wind speed (m/s)
	DVvvmin	Wind direction of VVmin (1 ~ 8)
	DVdom	Daily dominant wind direction (1 ~ 8)
	DVvvmax	Wind direction of VVmax (1 ~ 8)
	HRmin	Daily minimum relative humidity (%)
	HRmoy	Daily mean relative humidity (%)
	HRmax	Daily maximum relative humidity (%)
	SR	Daily mean solar radiation (W/m^2)
Other pollutants	SO_2	Daily average concentration of Sulfur dioxide
	NO_2	Daily average concentration of Nitrogen dioxide
	NO	Daily average concentration of Nitric oxide
	O_3	Daily average concentration of Ozone
Lagged PM_{10}	$PM_{10}(j - t)$	Daily average concentration of PM_{10} in day $j - t$ (t = 1, 2, ..., 7)

3.1. Data Gathering and Description

Various panniers of explicative variables have been used in the previous works for the purpose of PM_{10} modeling and forecasting (Dong et al., 2009; Kurt and Oktay, 2010; Poggi and Portier, 2011; Domańska and Wojtylak, 2012). This variety depends on the availability of measured variables and the objectives of the study. Meteorological fields, including wind, hourly temperature, mixing depth and solar insolation fields are an important input for any modeling exercise with air quality models. These fields can have great uncertainty which contribute in mispredicting airborne chemical species, aerosols and particulate matter. Thus, accurate meteorological fields are of utmost importance. They will lead to reliable forecasts of air pollution events (Almanza et al., 2014).

In our study we do not care about the number of used explicative variables and their contribution to explain the ozone variation as our main goal is to pick out the best statistically.

The dataset used in this study consists of PM_{10} daily average concentrations, other pollutants (SO_2, NO_2, NO and O_3) and meteorological data observed in two monitoring stations from Tunisia. The first station is installed at Gabes. Its database contains 349 observations from 01/01/2010 to 15/12/2010. The second station is at Manouba, and its database contains 154 observations from 11/05/2010 to 11/10/2010. The data 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.

The objective of this study is to model, to analyze and to forecast the PM_{10} daily average concentrations using the SVR and the RF approaches. To that effect, we use twenty-four explicative variables to predict the PM_{10} daily average concentration. These variables are grouped into three categories; meteorological indicators, other pollutants and delayed PM_{10} daily average concentrations. Table 1 summarizes all the explicative variables. In some previous studies, it was shown that the first lagged PM_{10} is an important predictor of its current value (at day j) but without identifying statistically the proper delay (Chaloulakou et al., 2003; Poggi and Portier, 2011). They are limited only to one lag of PM_{10}. Here we consider seven delayed PM_{10} concentrations from j - 1 to j - 7. The best predictors to keep in the model will be statistically identified hereafter.

We note that the variables associated with wind direction

DVdom, DVvvmin and DVvvmax 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 north-west = 8. This is the wind compass describing the eight principal bearings used habitually in meteorology to categorize wind direction.

Figure 2 shows the evolution of PM_{10} daily average concentrations (in µg/m³) for each monitoring station. As it can be seen, the two monitoring stations are different from the daily average PM_{10} concentrations variation. We note a large variability in the PM_{10} values for the two stations. Moreover, Manouba database contains more outliers than that of Gabes. This tricky behavior can cause, probably, some difficulties in its modeling and forecasting.

3.2. Missing Values Treatment

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 two collected raw databases present many missing values. For Gabes database, 6.99% of the data does not exist and for Manouba database 23.65% of the data are missed. To handle this problem of missing values, we use an imputation technique based on a multivariate imputation by chained equations developed by Van Buuren and Groothuis-Oudshoorn (2011) and implemented in the MICE algorithm on the software R freely downloadable from http://cran.r-project.org/.

4. Experimental Results

This section reports the experimental results on the above presented variable importance scores. This is done on simulated and on two real world datasets. The purpose of our empirical analysis is twofold. First, for simulated data, we check the ability of all the scores to properly rank the right important variables within a linear regression framework when noisy variables are added. Then we check the ability of our stepwise algorithm to select the best subset of variables. Second, for real world data, we compare the variable selection methods using datasets from the two stations and we focus on the forecasting performance improvement. We have used MATLAB for our computational tasks.

4.1. Guideline Results on Synthetic Data

The major advantage of using simulated data here is that the relevant variables are already known by construction. So, this part can give us some judgments and benchmark results that will be used as guidelines for the real application.

We consider a linear relation between the target y and the first six input variables x_i, i = 1, 2, …, 6 given by:

$$y = x_1 - 2x_2 + 3x_3 - 4x_4 + 5x_5 - 6x_6 \quad (4)$$

Only the first six variables are relevant and the added ones are noise. All the variables are randomly and independently drawn according to the standard normal distribution. Note that we can vary the number of noisy variables and the dataset size according to our aims.

As the target y is linearly related to the explicative variables we have used a standard linear SVR. To tune the other parameters of the SVR model, we have performed a grid search over several runs of 10-fold cross-validation which lead to choose $\varepsilon = 0.001$ and C = 1,000 whatever n and p.

For RF model, we used the results of Genuer et al. (2010) and Díaz-Uriarte and Alvarez de Andres (2006) and we made some preliminary simulations to find an optimal parameters tuning. According to the obtained results, parameters *nodesize* and *mtry* are set to their default values for regression (*nodesize* = 5 and *mtry* = p/3) but taking *ntree* = 300 leads to good stability. The performance gain was negligible in our simulations.

In order to check the ability of the five scores to retrieve the right pertinent variables, we have performed 100 trials with random draws of training set. At each trial we record the number of relevant variables correctly top ranked at the first six positions. This was done for different values of n and p.

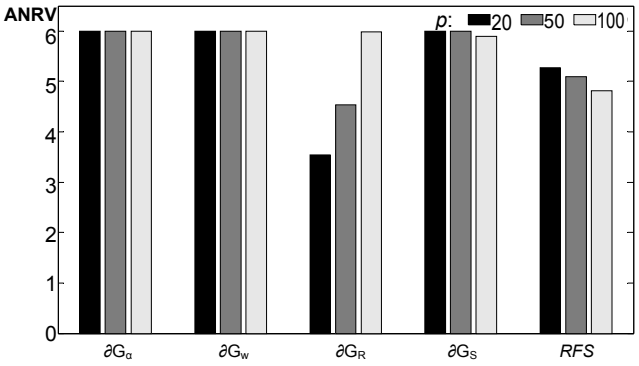

Figure 3. Average number of relevant variables (ANRV) correctly ranked with respect to the number of variables (p = 20, 50, 100) and the used score of importance. The sample size is set to n = 150.

4.1.1. Ranking Sensitivity to p

The experiment that we carried out here is the following. We set the sample size to n = 150 and we vary the number of variables taking p = 20, 50 and 100. All the five scores of importance have been computed as previously mentioned. The SVR scores are computed over 200 bootstrap samples. For RF, the importance score is averaged on several runs to avoid insignificant sampling effects. Each score gives rise to a decreasing hierarchy of importance. The number of relevant variables ranked in the top six positions of the hierarchies have been counted. For each value of p, the results have been averaged over 100 trials with random draws of training set. Figure 3 shows the average number of relevant variables ranked in the top six ranks when increasing the number of noisy variables.

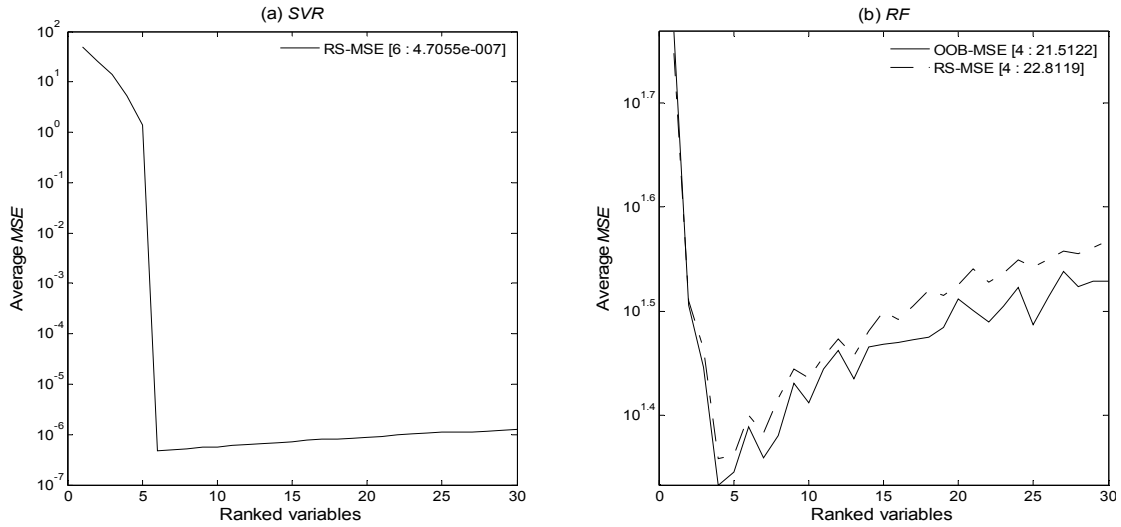

Figure 5. Mean Squared Error of nested increasing models on toy data. For each score, the minimum error rate and the corresponding optimal number of relevant variables are given in brackets. We take n = 150 and p = 30. The *y*-axis is taken in the logarithmic scale. (a) Averaged MSE over 50 random splitting for the $\partial G\alpha$ score; (b) Error rates estimated on *OOB* samples and over 50 random splitting for the RF score.

From Figure 3 we see that the scores ∂G_α, ∂G_W and ∂G_S outperform greatly the other scores in variables importance assessment in all situations. Moreover, an unexpected result is that the ∂G_R score performs better as soon as the number of variables becomes large enough. However, the effectiveness of the RF*S* score deteriorates by increasing the number of variables.

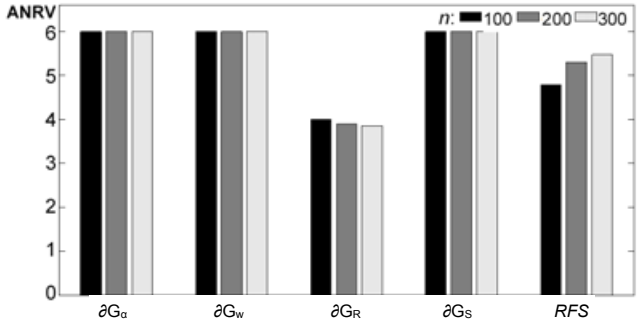

Figure 4. Average number of relevant variables (ANRV) correctly ranked with respect to the sample size (n = 100, 200, 300) and the used score of importance. The number of variables is set to p = 30.

4.1.2. Ranking Sensitivity to *n*

Now we fix the number of variables to p = 30 and we vary the sample size taking n = 100, 200 and 300. Then we do the same computations as previously. Figure 4 gives the average number of relevant variables ranked in the top six ranks when increasing the sample size.

The results presented in this figure confirm the previous findings. It seems that the ∂G_R score is more robust to noisy variables when their number becomes larger relatively to the sample size.

4.1.3. Stepwise Curve Shape

Let us now run our stepwise algorithm on the toy dataset using only the scores ∂G_α with the SVR and the score RF*S* with the RF. A sequence of nested increasing models is constructed on each hierarchy. The Mean Squared Error (*MSE*) for the SVR model is averaged over 50 random splits; 80% for learning and the remaining for testing. For the RF model, the *MSE* is estimated in two ways; over *OOB* samples and then over 50 random splits. For uniformity reasons, we keep the same 50 random splits used for the SVR and RF models. Finally, the model realizing the lowest *MSE* is chosen as the model having the optimal subset of variables. We used here a toy dataset with n = 150 and p = 30. Our aim is to check the ability of our stepwise algorithm to select the best subset of variables with the different approaches.

Figure 5 depicts the performance of the nested increasing models where variables are introduced sequentially in a decreasing order of importance. We observe that the two RF curves (right panel) have very similar 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 variables ranking. This behavior was deeply analyzed in the work of Ghattas and Ben Ishak (2008) for binary classification and more recently in the work of Feki et al. (2012) for multiclass problems. Moreover, the *OOB* curve (*OOB-MSE*) seems a little bit optimistic than the *RS* curve (*RS-MSE*) obtained by random splitting for RF. The left panel shows that the expected curve shape is less remarkable with SVR. Indeed, the rising phase of the *MSE* is slower. This shows that the SVR fo-

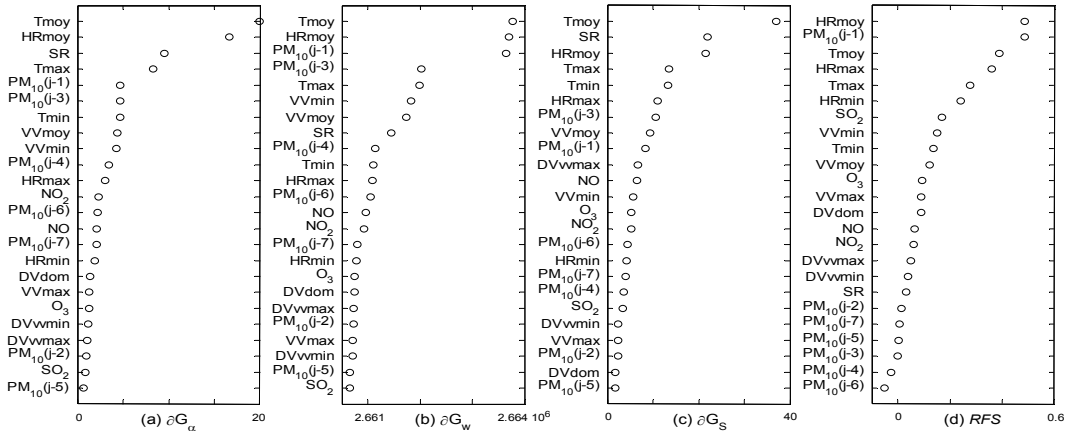

Figure 6. Variable ranking using the four scores of importance for Gabes station. (a) Variable ranking according to the score ∂G_α; (b) Variable ranking according to the score ∂G_W; (c) Variable ranking according to the score ∂G_S; (d) Variable ranking according to the score RF*S*.

recasting is less sensitive to the presence of noisy variables comparatively to RF model.

Although the simulation experiments deal with specific artificial situations, we can still draw useful and practical recommendations to properly conduct real applications. Thus, according to the previous results, in the real world applications we will use all the scores except ∂G_R which has the worst raking performance. Finally, we will propose a combined procedure between SVR and RF. It consists of using the SVR hierarchies to construct a sequence of nested increasing RF models and conversely. The goal of this combination is to gather the good qualities of the two models RF and SVR in a single procedure in hope to reduce selection bias effect (Ambroise and McLachlan, 2002). The problem of selection bias is especially remarkable with the SVR (left panel in Figure 5) where the average *MSE* reaches zero when only the six top ranked variables are introduced in the model.

Ultimately, for consistency reasons the error rates are estimated solely by random splitting for all choices and the results will be carefully compared and interpreted.

4.2. Real World Application

In this section we will present and compare the results obtained for the different approaches on the two considered stations Gabes and Manouba. 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 four scores of importance ∂G_α, ∂G_W, ∂G_S and RF*S* and then we select the subsets of relevant predictors using our stepwise algorithm. We will leave aside 10% of the observations chosen at random from each dataset for testing and selection bias checking.

At the beginning, we have performed a grid search over several runs of 10-fold cross-validation. The obtained results lead to take $\varepsilon = 0.001$ and $C = 1$ for the two datasets. The best kernel to use was polynomial with degree $d = 1$ meaning that the two datasets can be considered as linear.

Like in simulation part for RF model, parameters *nodesize* and *mtry* are set to their default values for regression (*nodesize* = 5 and *mtry* = p/3) and we took *ntree* = 300 which ensure good stability.

4.2.1. Experiments on Training Sets

All the work will be conducted here on the training sets. The random division gives rise to training sets for Gabes and Manouba stations containing 314 and 139 observations respectively. The remaining instances from each station are kept aside for testing and for selection bias checking.

The training sets are used to compute the variable importance according to the three retained SVR scores and the RF score. All the computations are carried out in a similar manner to that of the simulated part. Table 2 gives the Spearman's rank correlation coefficients ρ in order to measure the similarities between the different hierarchies across scores or/and stations.

Examining the Table 2 leads us to the following conclusions. The three SVR hierarchies are more similar to each other than that resulting from RF. This is true for the two stations, but the degree of similarity is a little higher for Gabes station. When we compare the similarities across stations we observe a significant difference between the hierarchies. This difference is much greater for the RF score.

Table 2. Spearman's Rank Correlation Coefficients for Comparison of the Hierarchies across Scores or/and Stations

		Gabes				Manouba			
		∂G_α	∂G_W	∂G_S	RF*S*	∂G_α	∂G_W	∂G_S	RF*S*
Gabes	∂G_α	1	0.96	0.84	0.39	0.23	0.15	0.38	-0.01
	∂G_W		1	0.82	0.38	0.22	0.18	0.31	0.04
	∂G_S			1	0.48	0.38	0.24	0.49	0.11
	RF*S*				1	0.59	0.48	0.59	-0.14
Manouba	∂G_α					1	0.90	0.80	0.31
	∂G_W						1	0.72	0.45
	∂G_S							1	0.28
	RF*S*								1

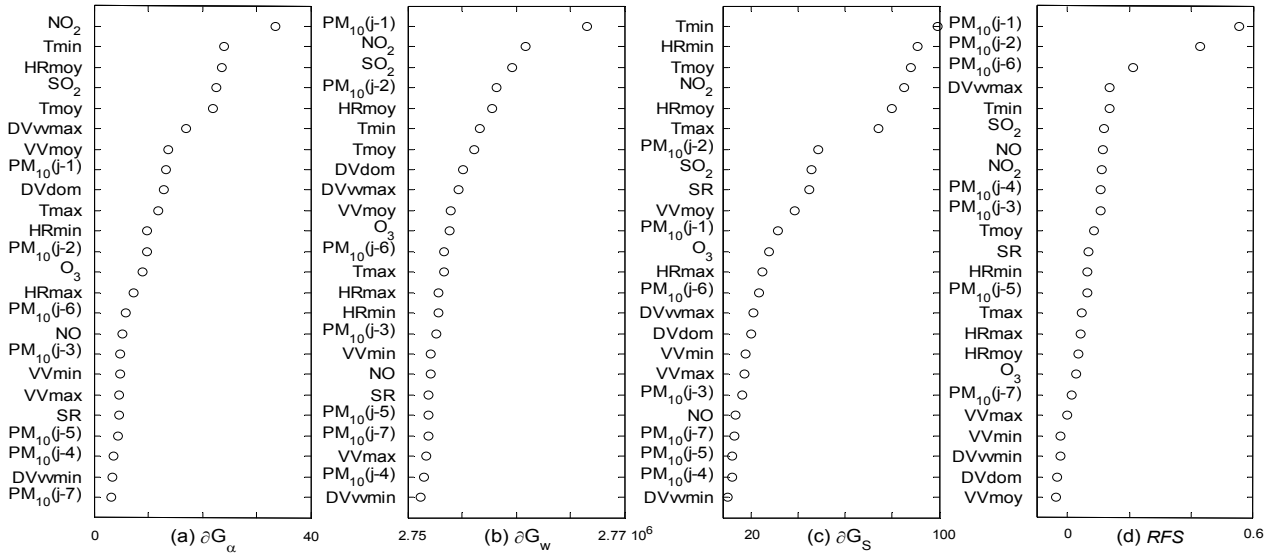

Figure 7. Variable ranking using the four scores of importance for Manouba station. (a) Variable ranking according to the score ∂G_α; (b) Variable ranking according to the score ∂G_W; (c) Variable ranking according to the score ∂G_S; (d) Variable ranking according to the score RF*S*.

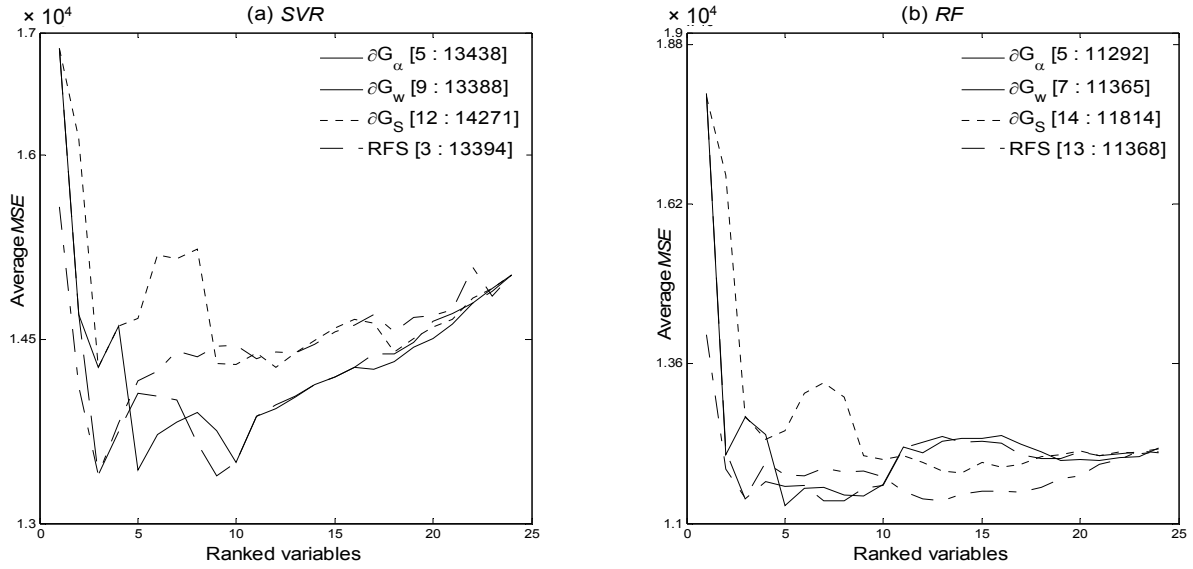

Figure 8. Gabes station: Mean Squared Error of nested increasing models. For each curve, the optimal number of relevant predictors and the corresponding *MSE* are given in brackets. (a) The nested SVR models; (b) The nested RF models.

To get a clearer idea about the different hierarchies, Figures 6 and 7 come to expose the variable ranking and the corresponding scores values for Gabes and Manouba stations, respectively.

From Figure 6 we notice a strong similarity between the headers of the three SVR hierarchies. Indeed, the first ten positions contain eight common variables. On the other hand, the RF*S* hierarchy is a little different. In fact, only six variables are common over the top ten ranks when we observe the four hierarchies simultaneously. These similarity statistics are lower for Manouba station. Moreover, we note that the top ranked variables are not the same from one station to the other which can be explained by their urban, meteorological and geographic differences. Finally, we can conclude that the variables related to temperature and relative humidity and the variable PM_{10} (j -1) are predominantly more or less top ranked whatever the score and the station. This finding is consistent with the results of the previous work of Poggi and Portier (2011).

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

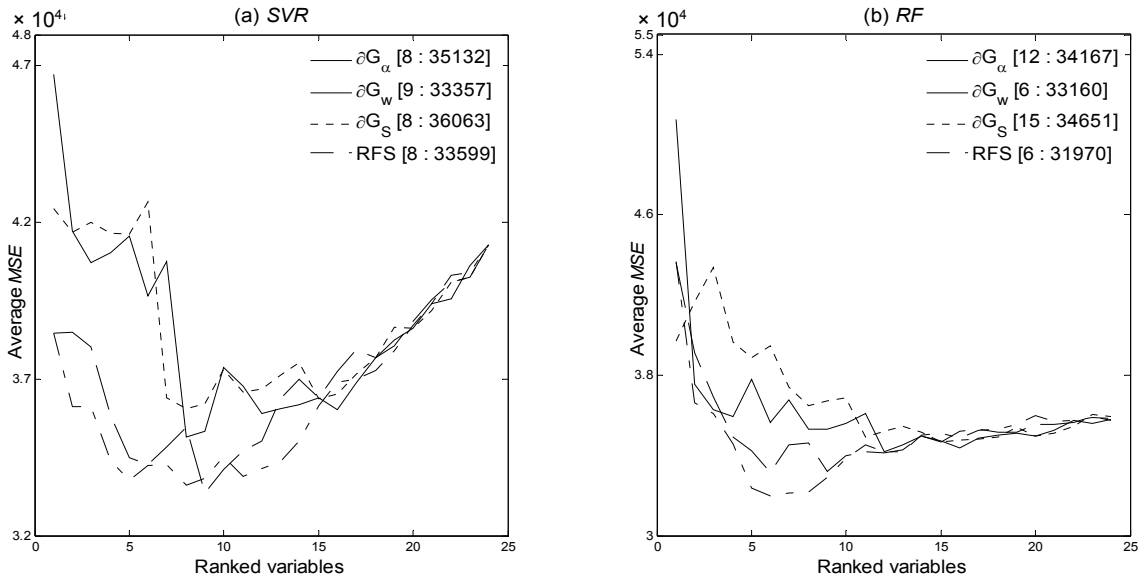

Figure 9. Manouba station: Mean Squared Error of nested increasing models. For each curve, the optimal number of relevant predictors and the corresponding *MSE* are given in brackets. (a) The nested SVR models; (b) The nested RF models.

our comparative study.

For the variable selection step, we will perform our stepwise algorithm with both RF and SVR using all the previous hierarchies. Using an external score to the model in the stepwise algorithm should reduce the selection bias problem (Ambroise & McLachlan, 2002). The optimal subset of predictors is the one achieving the lowest *MSE* over 50 random splitting; 80% for learning and 20% for testing. For homogeneity reasons, we have chosen to use random splitting instead of the *OOB* samples for the RF.

When we examine Figures 8 and 9 we directly note that the error rates are much higher than those in the simulated part (see Figure 5). These rates do not seem very strange and can be explained by the fact that the real datasets are not perfectly linear and on the boundary to be rather nonlinear. Besides, the dataset of Gabes station seems to have a more linear trend than that of Manouba station. We recall that the two real datasets are nevertheless linear according to our grid search testing. These results confirm again that real world datasets remain always much more complex than the simulated special cases. However, we cannot totally ignore the usefulness of simulation experiments.

What is more interesting is that all the curves depicted in Figures 8 and 9 show the expected typical behavior stressed in the simulated part in Figure 5. This typical curve shape is less respected when using the scores ∂G_W and ∂G_S. Moreover, the RF*S* hierarchy seems to be more suitable for the two datasets according to the corresponding *MSE* curve shape. This is due to the fact that RF are highly nonparametric and nonlinear learning models which fit well the data without overfitting especially when the data are nonlinear or faintly linear.

Finally, we see clearly that our variable selection approach improves significantly the forecasting performance by selecting a reduced numbers of predictors. The improvement magnitude and the subset size of selected predictors vary depending on the used score and model.

4.2.2. Selection Bias Checking

This paragraph is devoted to control the selection bias problem on the test sets. It is known that this problem is inherent to the tasks of variable selection (Ambroise and McLachlan, 2002). Let us first denote the previous selected subsets for Gabes station by:

$G^5_{\partial G_\alpha}$: 5 top ranked variables in the ∂G_α hierarchy selected in common by the SVR and the RF models,

$G^9_{\partial G_W}$: 9 top ranked variables in the ∂G_W hierarchy selected by the SVR model,

$G^7_{\partial G_W}$: 7 top ranked variables in the ∂G_W hierarchy selected by the RF model,

$G^{12}_{\partial G_S}$: 12 top ranked variables in the ∂G_S hierarchy selected by the SVR model,

$G^{14}_{\partial G_S}$: 14 top ranked variables in the ∂G_S hierarchy selected by the RF model,

$G^3_{\partial G_{RFS}}$: 3 top ranked variables in the RF*S* hierarchy selected by the SVR model,

$G^{13}_{\partial G_{RFS}}$: 13 top ranked variables in the RF*S* hierarchy selected by the RF model.

For Manouba station, they are denoted by:

$M^8_{\partial G_\alpha}$: 8 top ranked variables in the ∂G_α hierarchy selected by the SVR model,

$M^{12}_{\partial G_\alpha}$: 12 top ranked variables in the ∂G_α hierarchy selected by the RF model,

$M^9_{\partial G_W}$: 9 top ranked variables in the ∂G_W hierarchy selected by

the SVR model,
$M^{6}_{\partial G_W}$: 6 top ranked variables in the ∂G_W hierarchy selected by the RF model,
$M^{8}_{\partial G_S}$: 8 top ranked variables in the ∂G_S hierarchy selected by the SVR model,
$M^{15}_{\partial G_S}$: 15 top ranked variables in the ∂G_S hierarchy selected by the RF model,
$M^{8}_{\partial G_{RFS}}$: 8 top ranked variables in the RF*S* hierarchy selected by the SVR model, and
$M^{6}_{\partial G_{RFS}}$: 6 top ranked variables in the RF*S* hierarchy selected by the RF model.

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 (Legates and McCabe, 2013; Willmott et al., 2012; Krause et al., 2005). 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 (Wang et al., 2015; Antanasijević et al., 2013; Koo et al., 2012). 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:

$$RMSE = \sqrt{\frac{1}{l}\sum_{i=1}^{l}\left(O_i - P_i\right)^2} \tag{5}$$

$$MAE = \frac{1}{l}\sum_{i=1}^{l}\left|O_i - P_i\right| \tag{6}$$

$$MAPE = \frac{1}{l}\sum_{i=1}^{l}\left|\frac{O_i - P_i}{O_i}\right| \times 100\% \tag{7}$$

$$FA_2 = \frac{1}{l}\sum_{i=1}^{l}\chi_{[0.5,2]}\left(\frac{P_i}{O_i}\right) \tag{8}$$

$$FA_{1.25} = \frac{1}{l}\sum_{i=1}^{l}\chi_{[0.8,1.25]}\left(\frac{P_i}{O_i}\right) \tag{9}$$

$$d_r = \begin{cases} 1 - \dfrac{\sum_{i=1}^{l}\left|O_i - P_i\right|}{2\sum_{i=1}^{l}\left|O_i - \bar{\boldsymbol{O}}\right|}, \; if \; \sum_{i=1}^{l}\left|O_i - P_i\right| \le 2\sum_{i=1}^{l}\left|O_i - \bar{\boldsymbol{O}}\right| \\ \dfrac{\sum_{i=1}^{l}\left|O_i - P_i\right|}{2\sum_{i=1}^{l}\left|O_i - \bar{\boldsymbol{O}}\right|} - 1, \; otherwise \end{cases} \tag{10}$$

$$E_1 = 1 - \frac{\sum_{i=1}^{l}\left|O_i - P_i\right|}{\sum_{i=1}^{l}\left|O_i - \bar{\boldsymbol{O}}\right|} \tag{11}$$

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 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 (Willmott et al., 2012). Finally, to date, the E_1 coefficient is the main competitor with d_r (Legates and McCabe, 2013). For the last four metrics, the higher the value is, the better the quality of forecasts is.

Table 3. Gabes Station: Forecasting Accuracy Metrics for the Selected Subsets of Predictors and for all the Predictors

	RMSE	*MAE*	*MAPE*	FA_2	$FA_{1.25}$	d_r	E_1
SVR							
All variables	70.57	52.03	50.40%	0.83	0.31	0.49	-0.008
$G^{5}_{\partial G\alpha}$	73.74	51.30	45.82%	0.80	0.40	0.50	0.005
$G^{9}_{\partial Gw}$	74.01	55.71	53.94%	0.77	0.25	0.46	-0.08
$G^{12}_{\partial GS}$	72.50	51.89	49.58%	**0.85**	0.34	0.49	-0.005
$G^{3}_{\partial GRFS}$ (*)	**70.32**	**46.76**	**40.49%**	**0.85**	**0.43**	**0.54**	**0.09**
RF							
All variables	66.67	45.76	41.01%	0.85	0.47	0.55	0.11
$G^{5}_{\partial G\alpha}$	**61.40**	44.30	40.20%	0.86	0.43	0.57	0.14
$G^{7}_{\partial Gw}$	63.28	44.02	39.11%	**0.88**	0.43	0.57	0.15
$G^{14}_{\partial GS}$	65.51	46.02	41.19%	0.85	0.41	0.55	0.11
$G^{13}_{\partial GRFS}$ (*)	65.22	**43.47**	**38.27%**	**0.88**	**0.48**	**0.58**	**0.16**

Tables 3 and 4 give the predictive performance realized by the SVR and the RF models on the test sets when using all variables and when using only the selected subsets for the both

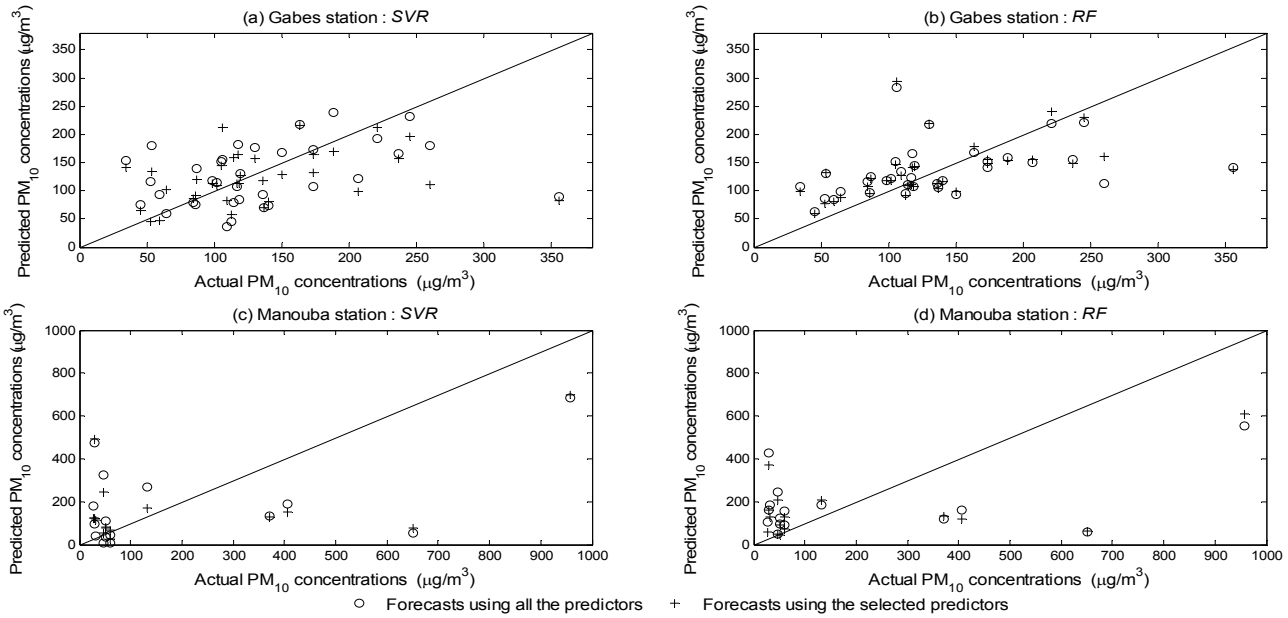

Figure 10. Actual values versus forecasts of PM_{10} concentrations for the test sets. Comparison between the forecasts using all the predictors and those using only the selected predictors. (a) Gabes station with the SVR model; (b) Gabes station with the RF model; (c) Manouba station with the SVR model; (d) Manouba station with the RF model.

stations. For RF model, we have reported the average error on several runs in order to attenuate the random sampling effects inherent to the RF architecture and to provide fairest results.

From Tables 3 and 4 we see that, in most cases, the variable selection improves significantly the forecasting accuracy on the test sets. This proves that our variable selection procedure is not seriously affected by the selection bias problem. The best result for each criterion is written in bold. The most best subset of predictors for each pair model/station is marked by an asterisk. It is the subset which improves the majority of the adopted criteria. The overall improvement, according to the seven criteria, is more important for Gabes station. For instance, the best relative gain in forecasts accuracy (in terms of *MAPE*) is about 19.66% ((50.40 - 40.49)/50.40) when using the SVR model with the subset of predictors $G^3_{\partial G_{RFS}}$, and is approximately 16.29% ((230.11 - 192.63)/230.11) when using the RF model with the subset of predictors $M^6_{\partial G_{RFS}}$ for Gabes and Manouba stations, respectively. Moreover, the best relative improvement in the *RMSE* is slightly greater for the RF model in favor of Gabes station. Indeed, the best relative gain in the *RMSE* is about 7.9% ((66.67 - 61.4)/66.67) when using the RF model with the subset of predictors $G^5_{\partial G_\alpha}$ for Gabes, and is approximately 7.03% ((24 3.62 - 226.49)/243.62) when using the RF model with the subset of predictors $G^{12}_{\partial G_\alpha}$ for Manouba. These results show that the selection bias is much more attenuated when using an external score to the involved model. 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.

Table 4. Manouba Station: Forecasting Accuracy Metrics for the Selected Subsets of Predictors and for all the Predictors

	RMSE	*MAE*	*MAPE*	FA_2	$FA_{1.25}$	d_r	E_1
SVR							
All variables	239.94	172.86	235.60%	0.26	0	0.59	0.19
$M^8_{\partial G\alpha}$	**226.41**	164.08	**211.12%**	**0.40**	**0.13**	0.61	0.23
$M^9_{\partial Gw}$ (*)	231.89	**160.60**	222.81%	**0.40**	**0.13**	**0.62**	**0.24**
$M^8_{\partial GS}$	241.82	179.24	265.01%	**0.40**	**0.13**	0.58	0.16
$M^8_{\partial GRFS}$	241.78	170.50	218.03%	**0.40**	**0.13**	0.60	0.19
RF							
All variables	243.62	178.75	230.11%	0.33	0.07	0.58	0.16
$M^{12}_{\partial G\alpha}$	**226.49**	161.71	205.58%	0.40	0.13	**0.62**	**0.24**
$M^6_{\partial Gw}$	232.07	**160.97**	207.50%	0.41	**0.19**	**0.62**	**0.24**
$M^{15}_{\partial GS}$	233.34	166.10	202.19%	0.43	0.07	0.61	0.22
$M^6_{\partial GRFS}$ (*)	229.27	162.32	**192.63%**	**0.44**	**0.19**	**0.62**	**0.24**

Finally, Figure 10 shows the forecasting performance of the selected subsets of predictors compared to using all the predictors. The observed values versus the model forecasts of PM_{10} concentrations are depicted for each dataset. Each scatter plot corresponds to one of the models marked by an asterisk in Tables 3 and 4. We can see that the overall quality of forecasts is at least preserved when using only the selected predictors.

5. Conclusions

In this work, we have compared two popular statistical learning models namely the Support Vector Regression and the Random Forests for the purpose of variable selection and fore-

casting. This comparison study was conducted on synthetic and real datasets. The results of the simulated part were used as a benchmark to properly conduct the real application. We have considered two monitoring stations from Tunisia to model and forecast the PM_{10} daily average concentration. The problem of variable selection for linear multiple regression was deeply investigated.

On the linear simulated data we have shown that the SVR scores ∂G_α , ∂G_W and ∂G_S outperform slightly the score RFS in variable importance assessment. It has been also demonstrated that the score ∂G_R is comparatively much less efficient. Concerning the real world application, we have noticed that the linear trend was not explicit in the two datasets. Besides, Manouba dataset modeling was a little trickier than that for Gabes dataset because of outliers. Despite this difficulty, we were able to substantially improve the accuracy of forecasts for the two datasets. To do this, we have proposed a combined variable selection approach using RF and SVR simultaneously. The best improvement in the *RMSE* for the two datasets was achieved by using the score ∂G_α for variable ranking and the nested RF models for subset selection. This result is not surprising given that our variable selection procedure is based on the training set's *MSE* minimization. We have also demonstrated that our combined approach does not suffer from the problem of selection bias. This was done by considering various metrics of forecasts accuracy.

In practice, we have shown that it is possible to accurately forecast the PM_{10} daily average concentration by using only a reduced number of selected variables. The number of selected variables differs from one station to the other. This variability can be explained by their urban, meteorological and geographic large differences. Nevertheless, we have identified four common variables namely Tmoy, Tmax, HRmoy and $PM_{10}(j - 1)$.

Of course, the problem of variable selection for regression remains one of the main open issues in statistics. This challenge is certainly more difficult when we deal with nonlinear regression and/or handle situation exposing the curse of dimensionality phenomenon with a lot of highly correlated variables like in microarray data. Finally, the scope of our application could be broadened to cover other monitoring stations and by considering supplementary explicative variables.

Acknowledgments. The author is extremely grateful to the editor and the anonymous reviewers for their insightful comments and valuable suggestions, which contributed much to improving the manuscript.

References

Almanza, V.H., Batyrshin, I., and Sosa, G. (2014). Multi-criteria selection for an Air Quality Model configuration based on quantitative and linguistic evaluations. *Expert Syst. Appl.,* 41(3), 869-876. http://dx.doi.org/10.1016/j.eswa.2013.08.017

Amaldi, E., and Kann, V. (1998). On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theor. Comput. Sci.,* 209(1-2), 237-260. http://dx.doi.org/10.1016/S0304-3975(97)00115-1

Ambroise, C., and McLachlan, G.J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. of the National Academy of Sciences of the United States of America*, 99(10), 6562-6566. http://dx.doi.org/10.1073/pnas.102102699

Antanasijević, D.Z., Pocajt, V.V., Povrenović, D.S., Ristić, M.Đ., and Perić-Grujić, A.A. (2013). PM_{10} emission forecasting using artificial neural networks and genetic algorithm input variable optimization. *Sci. Total Environ.,* 443, 511-519. http://dx.doi.org/10.1016/j.scitotenv.2012.10.110

Ben Ishak, A., and Ghattas, B. (2005). An efficient method for variable selection using svm-based criteria. Preprint IML, l'Institut de Mathématiques de Luminy, Marseille, France. Available at http://iml.univ-mrs.fr/editions/preprint2005/preprint2005.html.

Boser, B.E., Guyon, I.M., and Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. *Proc. of the Fifth Annual Workshop on Computational Learning Theory*, ACM, Pittsburgh, 144-152. http://dx.doi.org/10.1145/130385.130401

Breiman, L. (1996). Bagging predictors. *Mach. Learning,* 24(2), 123-140. http://dx.doi.org/10.1023/A:1018054314350

Breiman, L. (2001). Random Forests. *Mach. Learning,* 45(1), 5-32. http://dx.doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone C.J. (1984). *Classification and Regression Trees*, Wadsworth and Brooks.

Carnevale, C., Finzi, G., Pisoni, E., Singh, V., and Volta, M. (2011). An integrated air quality forecast system for a metropolitan area. *J. Environ. Monit.,* 13, 3437-3447. http://dx.doi.org/10.1039/c1em10303b

Chaloulakou, A., Saisana, M., and Spyrellis, N. (2003). Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens. *Sci. Total Environ.,* 313(1-3), 1-13. http://dx.doi.org/10.1016/S0048-9697(03)00335-8

Chang, M.W., and Lin, C.J. (2005). Leave-one-out bounds for support vector regression model selection. *Neural Computation,* 17(5), 1188-1222. http://dx.doi.org/10.1162/0899766053491869

Corani, G. (2005). Air quality prediction in Milan: Feed forward neural networks, pruned neural networks and lazy learning. *Ecol. Model.,* 185(2-4), 513-529. http://dx.doi.org/10.1016/j.ecolmodel.2005.01.008

Cordelino, C., Chang, M., St John, J., Murphey, B., Cordle, J., Ballagas, R., Patterson, L., Powell, K., Stogner, J., and Zimmer-Dauphinee, S. (2001). Ozone prediction in Atlanta Georgia: Analysis of the 1999 ozone season. *J. Air Waste Manage. Assoc.,* 51(8), 1227-1236. http://dx.doi.org/10.1080/10473289.2001.10464342

Cristianini, N., and Taylor, J.S. (2000). *An Introduction to Support Vector Machines and Other Kernel Based Learning Methods*, Cambridge University Press. http://dx.doi.org/10.1017/cbo9780511801389

Díaz-Uriarte, R., and Alvarez de Andrés S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinform.,* 7(3). http://dx.doi.org/10.1186/1471-2105-7-3

Domańska, D., and Wojtylak, M. (2012). Application of fuzzy time series models for forecasting pollution concentrations. *Expert Syst. Appl.,* 39(9), 7673-7679. http://dx.doi.org/10.1016/j.eswa.2012.01.023

Dong, M., Yang, D., Kuang, Y., He, D., Erdal, S., and Kenski, D. (2009). $PM_{2.5}$ concentration prediction using hidden semi-Markov model-based times series data mining. *Expert Syst. Appl.,* 36(5), 9046-9055. http://dx.doi.org/10.1016/j.eswa.2008.12.017

Feki, A., Ben Ishak, A., and Feki, S. (2012). Feature selection using Bayesian and multiclass Support Vector Machines approaches: Application to bank risk prediction. *Expert Syst. Appl.,* 39(3), 3087-3099. http://dx.doi.org/10.1016/j.eswa.2011.08.172

Genuer, R., Poggi, J.M., and Tuleau-Malot, C. (2010). Variable

selection using random forests. *Pattern Recognition Lett.,* 31(14), 2225-2236. http://dx.doi.org/10.1016/j.patrec.2010.03.014

Ghattas, B., and Ben Ishak, A. (2008). Sélection de variables pour la classification binaire en grande dimension: Comparaisons et application aux données de biopuces. *J. Soc. Fr. Stat. Rev. Stat. Appl.,* 149(3), 43-66.

Gregorutti, B., Michel, B., and Saint-Pierre, P. (2015). Grouped variable importance with random forests and application to multivariate functional data analysis. *Comput. Stat. Data Anal.,* 90, 15-35. http://dx.doi.org/10.1016/j.csda.2015.04.002

Grivas, G., and Chaloulakou, A. (2006). Artificial neural network models for prediction of PM_{10} hourly concentrations, in the Greater Area of Athens, Greece. *Atmos. Environ.,* 40(7), 1216-1229. http://dx.doi.org/10.1016/j.atmosenv.2005.10.036

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learning Res.,* 3, 1157-1182.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Mach. Learning,* 46(1-3), 389-422. http://dx.doi.org/10.1023/A:1012487302797

Hauck, H., Berner, A., Frischer, T., Gomiscek, B., Kundi, M., Neuberger, M., Puxbaum, H., and Preining, O. (2004). AUPHEP -- Austrian project on health effects of particulates -- general overview. *Atmos. Environ.,* 38(24), 3905-3915. http://dx.doi.org/10.1016/j.atmosenv.2003.09.080

Hoi, K.I., Yuen, K.V., and Mok, K.M. (2009). Prediction of daily averaged PM_{10} concentrations by statistical time-varying model. *Atmos. Environ.,* 43(16), 2579-2581. http://dx.doi.org/10.1016/j.atmosenv.2009.02.020

Huang, Q., Cheng, S.Y., Li, Y.P., Li, J.B., Chen, D.S., and Wang, H.Y. (2010). An integrated MM5-CAMx modeling approach for assessing PM_{10} contribution from different sources in Beijing, China. *J. Environ. Inf.,* 15(2), 47-61. http://dx.doi.org/10.3808/jei.201000166

Koo, Y.S., Kim, S.T., Cho, J.S., and Jang, Y.K. (2012). Performance evaluation of the updated air quality forecasting system for Seoul predicting PM_{10}. *Atmos. Environ.,* 58, 56-69. http://dx.doi.org/10.1016/j.atmosenv.2012.02.004

Krause, P., Boyle, D.P., and Bäse, F. (2005). Comparison of different efficiency criteria for hydrological model assessment. *Adv. Geosci.,* 5, 89-97. http://dx.doi.org/10.5194/adgeo-5-89-2005

Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., Niska, H., Dorling, S., Chatterton, T., Foxall, R., and Cawley, G. (2003). Extensive evaluation of neural network models for the prediction of NO_2 and PM_{10} concentrations, compared with a deterministic modeling system and measurements in central Helsinki. *Atmos. Environ.,* 37(32), 4539-4550. http://dx.doi.org/10.1016/S1352-2310(03)00583-1

Kurt, A., and Oktay, A.B. (2010). Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks. *Expert Syst. Appl.,* 37(12), 7986-7992. http://dx.doi.org/10.1016/j.eswa.2010.05.093

Legates, D.R., and McCabe, G.J. (2013). A refined index of model performance: A rejoinder. *Int. J. Climatol.,* 33(4), 1053-1056. http://dx.doi.org/10.1002/joc.3487

Moshammer, H., and Neuberger, M. (2003). The active surface of suspended particles as a predictor of lung function and pulmonary symptoms in Austrian school children. *Atmos. Environ.,* 37(13), 1737-1744. http://dx.doi.org/10.1016/S1352-2310(03)00073-6

Ortiz-García, E.G., Salcedo-Sanz, S., Pérez-Bellido, Á.M., Portilla-Figueras J.A., and Prieto, L. (2010). Prediction of hourly O_3 concentrations using support vector regression algorithms. *Atmos. Environ.,* 44(35), 4481-4488. http://dx.doi.org/10.1016/j.atmosenv.2010.07.024

Paschalidou, A.K., Kassomenos, P.A., and Bartzokas, A. (2009). A comparative study on various statistical techniques predicting ozone concentrations: Implications to environmental management. *Environ. Monit. Assess.,* 148(1-4), 277-89. http://dx.doi.org/10.1007/s10661-008-0158-0

Perez, L., Medina-Ramon, M., Konzli, N., Alastuey, A., Pey, J., Perez, N., Garcia, R., Tobias, A., Querol, X., and Sunyer, J. (2009). Size fractionate particulate matter, vehicle traffic, and case-specific daily mortality in Barcelona, Spain. *Environ. Sci. Technol.,* 43(13), 4707-4714. http://dx.doi.org/10.1021/es8031488

Poggi, J.M., and Portier, B. (2011). PM_{10} forecasting using clusterwise regression. *Atmos. Environ.,* 45(38), 7005-7014. http://dx.doi.org/10.1016/j.atmosenv.2011.09.016

Pope III, C.A., and Dockery D.W. (2006). Health effects of fine particulate air pollution: Lines that connect. *J. Air Waste Manage. Assoc.,* 56(6), 709-742. http://dx.doi.org/10.1080/10473289.2006.10464485

Pope III, C.A. (2000). Review: Epidemiological basis for particulate air pollution health standards. *Aerosol. Sci. Technol.,* 32(1), 4-14. http://dx.doi.org/10.1080/027868200303885

Qin, S., Liu, F., Wang, J., and Sun, B. (2014). Analysis and forecasting of the particulate matter (PM) concentration levels over four major cities of China using hybrid models. *Atmos. Environ.,* 98, 665-675. http://dx.doi.org/10.1016/j.atmosenv.2014.09.046

Rakotomamonjy, A. (2003). Variable selection using SVM-based criteria. *J. Mach. Learning Res.,* 3, 1357-1370.

Rakotomamonjy, A. (2007). Analysis of SVM regression bounds for variable ranking. *Neurocomputing,* 70(7-9), 1489-1501. http://dx.doi.org/10.1016/j.neucom.2006.03.016

Russell, A.G., and Brunekreef, B. (2009). A focus on particulate matter and health. *Environ. Sci. Technol.,* 43(13), 4620-4625. http://dx.doi.org/10.1021/es9005459

Sfetsos, A., and Vlachogiannis, D. (2010). Time series forecasting of hourly PM_{10} using localized linear models. *J. Softw. Eng. Appl.,* 3, 374-383. http://dx.doi.org/10.4236/jsea.2010.34042

Shawe-Taylor, J., and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*, Cambridge University Press, Cambridge, UK. http://dx.doi.org/10.1017/cbo9780511809682

Slini, T., Kaprara, A., Karatzas, K., and Moussiopoulos, N. (2006). PM_{10} forecasting for Thessaloniki, Greece. *Environ. Model. Software,* 21(4), 559-565. http://dx.doi.org/10.1016/j.envsoft.2004.06.011

Smola, A.J., and Schölkopf, B. (1998). *A Tutorial on Support Vector Regression*, NeuroCOLT Technical Report NC-TR-98-030, Royal Holloway College, University of London, UK.

Stadlober, E., Hörmann, S., and Pfeiler, B. (2008). Quality and performance of a PM_{10} daily forecasting model. *Atmos. Environ.,* 42(6), 1098-1109. http://dx.doi.org/10.1016/j.atmosenv.2007.10.073

Van Buuren, S., and Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *J. Stat. Software,* 45(3).

Vapnik, V., and Chapelle, O. (2000). Bounds on error expectation for support vector machines. *Neural Computation,* 12(9), 2013-2036. http://dx.doi.org/10.1162/089976600300015042

Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory*, Springer Verlag, New York. http://dx.doi.org/10.1007/978-1-4757-2440-0

Vapnik, V.N. (1998). *Statistical Learning Theory*, Wiley, New York.

Vapnik, V.N., Golowich, S., and Smola, A. (1997). Support vector method for function approximation regression estimation and signal processing, in M. Mozer, M. Jordan and T. Petsche (Eds.), *Advances in Neural Information Processing Systems pages*, Cambridge, MA, MIT Press.

Wang, P., Liu, Y., Qin, Z., and Zhang, G. (2015). A novel hybrid forecasting model for PM_{10} and SO_2 daily concentrations. *Sci. Total Environ.,* 505, 1202-1212. http://dx.doi.org/10.1016/j.scitotenv.2014.10.078

Willmott, C.J., Robeson, S.M., and Matsuura, K. (2012). A refined index of model performance. *Int. J. Climatol.,* 32(13), 2088-2094. http://dx.doi.org/10.1002/joc.2419

Yang, Z.C. (2014). Modeling and forecasting daily movement of am bient air mean $PM_{2.5}$ concentration based on the elliptic orbit model with weekly quasi-periodic extension: A case study. *Environ. Sci. Pollut. Res.,* 21(16), 9959-9972. http://dx.doi.org/10.1007/s11356-014-2899-3

Zhou, Y., Cheng, S.Y., Liu, L., and Chen, D.S. (2012). A Coupled MM5-CMAQ Modeling System for Assessing Effects of Restric tion Measures on PM_{10} Pollution in Olympic City of Beijing, China. *J. Environ. Inf.,* 19(2), 120-127.