PREDICTION OF MAXIMUM OZONE CONCENTRATION USING BIG DATA MODELS

KAID, Z. – ATTOUCH, M. – MASTEFAOUI, Z. – LAKSACI, A.*

Department of Mathematics, College of Science, King Khalid University
61413 Abha, Kingdom of Saudi Arabia

**Corresponding author*
e-mail: alikfa@kku.edu.sa

(Received 1st Jun 2019; accepted 10th Sep 2019)

Abstract. The air pollution problem which has arisen in developed countries, becoming evident by high levels of smoke from industries or traffic, has forced authorities to search for mechanisms to control the air quality by the real-time monitoring system. For this purpose, in this paper we develop a new procedure able to analyze this real-time data. More precisely, we use the recent development on mathematical Statistics to analyze the relationship between the maximum ozone concentration and the other palling gases such as the Nitric Oxides (NO), Nitrogen Dioxide (NO_2) and Sulphur Dioxide (SO_2). Specifically, we propose three models which are, Functional Nonparametric Regression, Functional Robust Regression and Functional Relative Error Regression. Considering the daily- curve of the concentration of the previous gases collected by the Marylebone road monitoring site in London, we provide statistical models allowing the prediction of the maximum ozone concentration 4 h ahead. We show that the accuracy of our prediction approaches is closely linked to the choice of the regression model and the input variables or the covariates. In particular, the nonparametric regression is more performant than the other models when the regressors are NO_2 and SO_2.

Keywords: *ozone forecasting, air quality data, functional regression, the nitric oxide, nitrogen dioxide, sulphur dioxide, nonparametric statistics, time series analysis, neural network model, Bayesian network models, principal component regression*

Introduction

Analyzing the levels of air pollution is nowadays of primary importance. Indeed the air quality has a great impact on human health as well as natural resources. In the last decades, air pollution is among the leading causes of death. Thus, pollution forecasting allows decision-makers to plan the prevention strategy. For this issue, several mathematical models are used to provide some software to control air quality (see, e.g., Ryan et al., 1995; Slini et al., 2002; Nghiem et al., 2009). Specifically, there exist two kinds of approaches: deterministic and statistical algorithms. The literature on the deterministic models treat the pollutant-formation processes by studying their chemical and physical properties (see Zhang et al., 2012). While the statistical models are based on the historical measurements of air pollution and/or meteorological data. In this context, most statistical models are constructed by using the classical regression method. We cite for instance the neural network models by Yi and Prybutok (1996), Bayesian network models by Gavrila (2013, 2016). We refer the readers to Gong and Ordieres-Meré (2016), Taylan (2017) and Ding et al. (2016) for some recent contributions on the prediction of ozone concentrations by the empirical methods. All the statistical models of these cited works are performed over the observed data in some discrete grid which allows at most making the daily prediction. However, because of the ozone concentration has a faster dynamics the daily prediction is not very important. Furthermore, the recent technological development of the measuring instruments and the informatics tools allow the recovery of increasingly bulky data being recorded densely over a thinner discretization grid what make them intrinsically a continuous curve. The manipulation of this kind of data permits the real-time forecasting of the ozone concentration. This is the main purpose of this paper. To do that, we use some new statistical models recently developed of big data analysis. The main advantage of these new models, so-called functional models, is the fact that they take into account the daily- curves in its continuous path, unlike the old models which take only the values of some few hours. Of course, with the functional models, we keep all the information in the sampled data and we predict different horizons. It should be noted that the functional statistics has encountered a strong infatuation in these last years, as evidenced the several special issues dedicated to this topic (see, e.g., Aneiros et al., 2019; Ferraty, 2010). For more discussion on this topic, we refer the reader to Hsing and Eubank (2015), Ling and Vieu (2018) and the references therein.

The present contribution deals with the assessment of functional statistics models for modelling ozone concentrations. More precisely, our main goal is to search accurate prediction methods of the ozone concentration with respect to the other polluting gases. We use for this study a sample of data recorded in the Marylebone road monitoring site in London. It contains the hourly measurements during the 2018-year for the following four variables: Nitric Oxide, Nitrogen Dioxide, Sulphur Dioxide and Ozone. This data is used to predict the maximum ozone concentration 4 h ahead. Let us point out that the considered models come from the recent development of the modern statistics that allow analyzing the big data without reduction of the dimension. These new approaches constitute alternative statistical models to the artificial neural network regression and the principal component regression which are usually employed in this prediction setting. The main feature of our functional models is the possibility to model the environmental date recorded by real-time monitoring. More precisely, we test three regression models that are Functional Nonparametric Regression (FNR), Functional Robust Regression (FRR) and the Functional Relative Regression (FRER) to forecast the maximum ozone concentration given four functional covariates such that the daily curves (one day before) of NO, NO_2, SO_2 and O_3. These models provide predictions, robust, fast and of higher accuracy.

The paper is organized as follows. In the next section, we introduce the functional methodology in the continuous-time prediction problem. The statistical models used are described in the section intituled “Some recent regression models”. The predictions results of the three proposed methods are gathered in result’s section. The last section is devoted to some conclusions.

Materials and methods

As mentioned in the previous section, the data of this contribution are acquired from real-time measurement by Marylebone road monitoring site. Marylebone Road is an important thoroughfare in central London, within the City of Westminster. It runs north-east to the south-west from the Euston Road at Regent’s park to the A40 Westway at Paddington. This road is frequently heavily congested. It carries about 90,000 vehicles a day. The Marylebone Road monitoring site is funded by the Department for Environment, Food and Rural Affairs (Defra) in June 1997. The monitoring cabin is located one meter from the kerb on the southern side of the road. Its geographical location are 51.522530 (latitude), -0.154611 (longitude). *Figure 1* shows the location of the site.

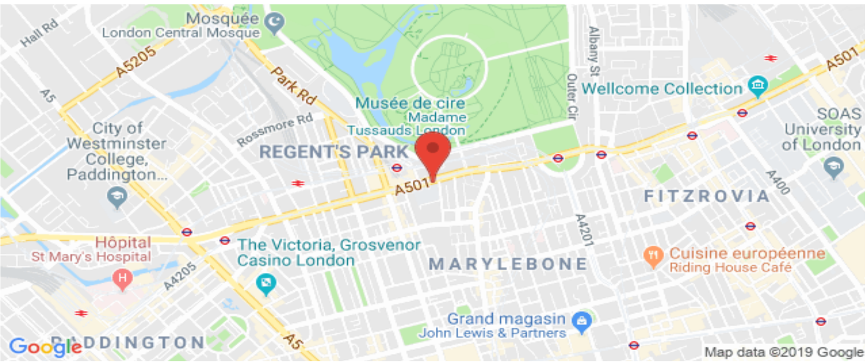

Figure 1. *The location of the Marylebone Road*

The data used in this paper are provided by the website https://www.airqualityengland.co.uk/site/data?site_id = MY1. It contains the hourly measurements during the period from January 1st to the 31st December for the year 2018, for the variables NO, NO_2, SO_2 and O_3 (*Fig. 2*).

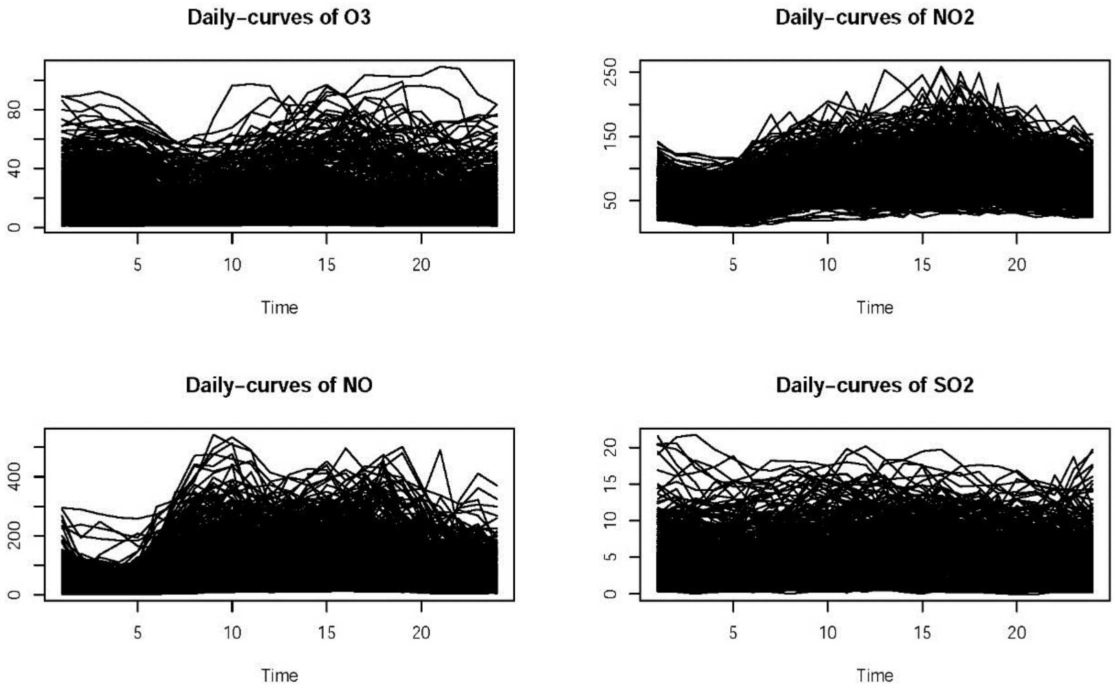

Figure 2. *The curves of the daily emission of the four gases NO, NO_2, SO_2 and O_3 in $\mu g/m^3$*

To fix the ideas, let us present the mathematical formulation of our prediction problem. Indeed, assume that we aim to predict the k-h-ahead prediction of the maximum ozone concentration at hour h_0, denoted by Y, using the curve of the daily emission of the gases observed the day before until $h_0 - k$. Formally, we assume that the output variable Y and the input variables $Z = (X_{NO}, X_{NO_2}, X_{SO_2}, X_{O_3})$ are linked by the following regression formula:

$$Y = r(Z) + error. \tag{Eq.1}$$

So, the prediction of Y is based on the determination of the function $r(.)$ in *Equation 1*. Many models are recently developed in mathematical statistics to resolve this kind of big data problem where the data are continuous curves. In this paper, we employ three regression procedures which are described in detail in the next section.

Some recent regression models

Functional nonparametric regression (FNR)

The nonparametric estimation of the functional regression was initially studied by Ferraty and Vieu (2006) and Ferraty et al. (2010). They used the Nadaraya Watson method to estimate this statistical model. Precisely, the function $r(.)$ is explicitly expressed using the least square error criterion by

$$r(x) = argmin_f E[(Y-f)^2|Z=z]. \tag{Eq.2}$$

It follows that $r(x) = E[Y|X=z]$.

So, for all fixed curves $x_{NO}, x_{NO_2}, x_{SO_2}, x_{O_3}$ we predict the maximum ozone concentration with respect to the criterion (*Eq. 2*) by

$$\hat{Y} = \hat{r}(z), \text{ where } z = (x_{NO}, x_{NO_2}, x_{SO_2}, x_{O_3})$$

where $\hat{r}(z)$ is the kernel estimator of $r(z)$ defined by

$$\hat{r}(x_{NO}, x_{NO_2}, x_{SO_2}, x_{O_3}) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{\|x_{NO}-x_{NOi}\|+\|x_{NO_2}-x_{NO2i}\|+\|x_{SO_2}-x_{SO2i}\|+\|x_{O_3}-x_{O3i}\|}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{\|x_{NO}-x_{NOi}\|+\|x_{NO_2}-x_{NO2i}\|+\|x_{SO_2}-x_{SO2i}\|+\|x_{O_3}-x_{O3i}\|}{h_n}\right)}$$

with K is a kernel function and h_n is a nonnegative real sequence.

Functional ρ- regression (FRR)

This regression model is obtained by resolving the following optimization problem

$$min_f E[\rho(Y,f)|Z=z], \tag{Eq.3}$$

ρ is a real-valued Borel function chosen by the user according to the studied data.

The model (*Eq. 3*) has been introduced in functional statistics by Azzedine et al. (2008) independent case and Attouch et al. (2012) for dependent case. The robustness is the main advantage of this model. It permits to analyze the data even in the presence of the outliers. Its functional estimation is expressed by

$$\bar{\bar{r}}\left(x_{NO},x_{NO_2},x_{SO_2},x_{O_3}\right) = argmin_f \frac{\sum_{i=1}^{n} \rho(Y_i,f) K\left(\frac{\|x_{NO}-x_{NOi}\|+\|x_{NO_2}-x_{NO2i}\|+\|x_{SO_2}-x_{SO2i}\|+\|x_{O_3}-x_{O3i}\|}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{\|x_{NO}-x_{NOi}\|+\|x_{NO_2}-x_{NO2i}\|+\|x_{SO_2}-x_{SO2i}\|+\|x_{O_3}-x_{O3i}\|}{h_n}\right)}.$$

Functional relative error regression (FRER)

This last regression is an alternative nonparametric regression to the least square regression model. It is recently considered in functional statistics by Demongeot et al. (2016) for independent case and Bassoudi and Kaid (2019) for dependent case. It is defined by the following rule

$$min_f\, E\left[\frac{(Y-f)^2}{Y^2} | Z = x\right]. \qquad \text{(Eq.4)}$$

The expression of this regression is explicitly given by

$$\frac{E[Y^{-1}|Z=z]}{E[Y^{-2}|Z=z]}.$$

Its estimator is defined by

$$\ddot{r}\left(x_{NO},x_{NO_2},x_{SO_2},x_{O_3}\right) = \frac{\sum_{i=1}^{n} Y_i^{-1} K\left(\frac{\|x_{NO}-x_{NOi}\|+\|x_{NO_2}-x_{NO2i}\|+\|x_{SO_2}-x_{SO2i}\|+\|x_{O_3}-x_{O3i}\|}{h_n}\right)}{\sum_{i=1}^{n} Y_i^{-2} K\left(\frac{\|x_{NO}-x_{NOi}\|+\|x_{NO_2}-x_{NO2i}\|+\|x_{SO_2}-x_{SO2i}\|+\|x_{O_3}-x_{O3i}\|}{h_n}\right)}.$$

Results

Undoubtedly, the accuracy of the three proposed regression models is closely related to the choice of the different parameters involved in these models. Specifically, the leading parameters in this prediction issue are kernel K, the smoothing parameter h_n, and the norm $\| \;\; \|$. The letter is equal to the distance between the smoothed curves obtained by interpolation. For basic materials on this notion, we refer the readers to Ferraty and Vieu (2006). Notice that the previous estimators are computed by using a sample of 364 obviations $(Y_i, X_{NOi}, X_{NO_{2i}}, X_{SO_{2i}}, X_{O_{3i}})$. The observations are determined with respect to the aimed prediction horizon. Generally speaking, for k-h-ahead prediction of O_3 at hour h_0 in a day i we denote by $\left(X_{NOi}, X_{NO_{2i}}, X_{SO_{2i}}, X_{O_{3i}}\right)$ the daily curve of the emission of the gases observed the day before $(i.e.\ (i-1))$ until $h_0 - k$, and we put $Y_i = X_{O_{3\,(i+1)}}(h_0)$. Now, to test the performance of the proposed models, we randomly split the 364 observations into two sub-samples: 200 in learning sample (indexed by i) and 164 in the testing sample (indexed by j). The observations of the learning sample are used to compute the estimators. While the observations of the test sample are employed to evaluate the quality of the models $(\tilde{r}, \bar{\bar{r}}, \ddot{r})$. To the end, let us point out that we have used the leave-one-out cross-validation technique to determine the bandwidth parameter h_n. The rule of the used cross-validation procedure is defined by

$$\frac{1}{200}\sum_{i=1}^{200}\left(Y_i - \theta(X_i)\right)^2 \qquad \text{(Eq.5)}$$

DOI: http://dx.doi.org/10.15666/aeer/1706_1423114239

where $\theta(.)$ means one of the previous regression models ($\hat{r}, \bar{\bar{r}}, \ddot{r}$). For the sake of shortness, we fixe $k = 4$ and we give the prediction results for $h_0 = 11\ am$. This hour corresponds to four hours after peak-hour road traffic. For the computational purpose, we use the same kernel and the score function as in Demongeot et al. (2016). To examine the sensitivity of this prediction problem to the input variables we have proceeded with several cases with different covariates. The obtained results are shown in *Figures 3–5*.

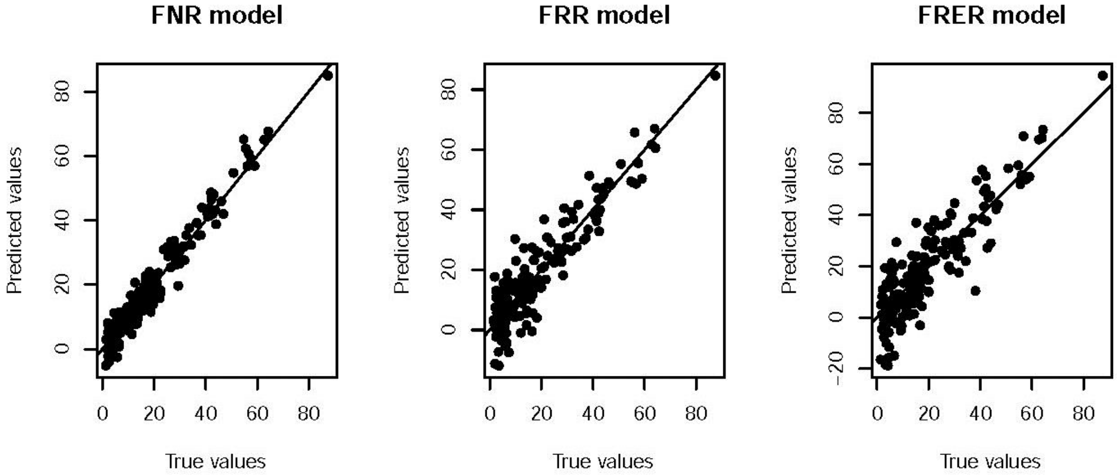

Figure 3. *4-h prediction of the maximum ozone concentration at $h_0 = 11\ am$ using the covariates $X_{NO}, X_{NO_2}, X_{O_3}$*

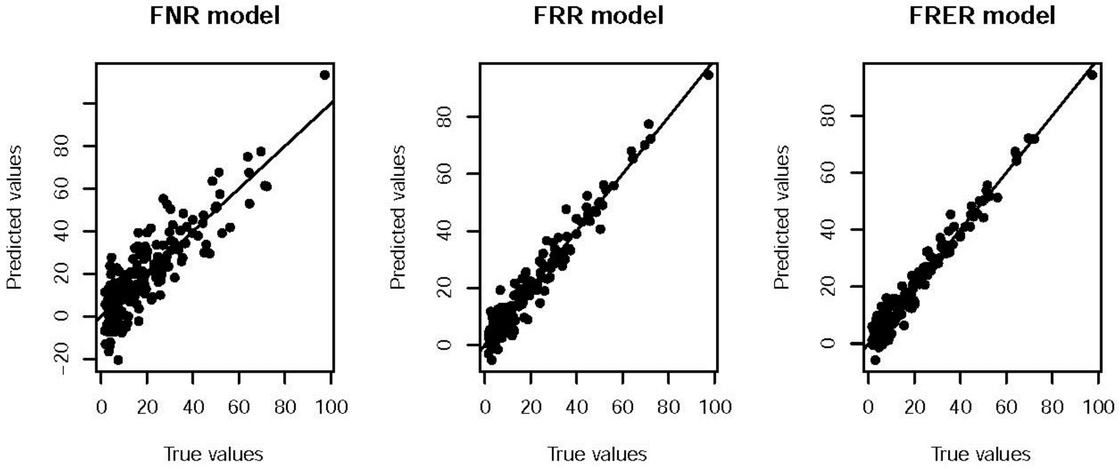

Figure 4. *4-h prediction of the maximum ozone concentration at $h_0 = 11\ am$ using the covariates X_{SO_2}, X_{O_3}*

DOI: http://dx.doi.org/10.15666/aeer/1706_1423114239

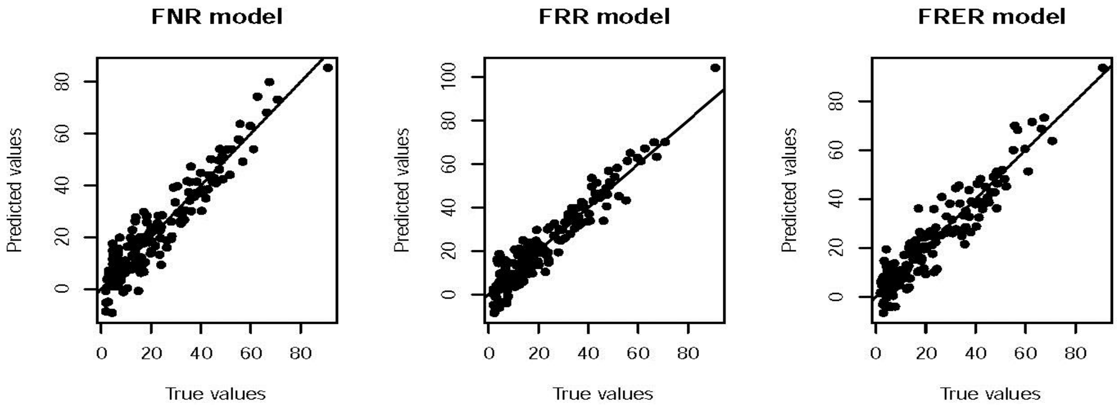

***Figure 5.** 4-h prediction of the maximum ozone concentration at $h_0 = 11\ am$ using the covariates $X_{NO}, X_{NO_2}, X_{SO_2}, X_{O_3}$*

It appears clearly that these new regression models performed well for this prediction problem and they have a fair accuracy. However, the performance of the considered models is varied with respect to the input variables. For instance, the FNR is more accurate when the regressors are $X_{NO}, X_{NO_2}, X_{O_3}$. On the other hand, the FRR and FRER provide better results for the covariates X_{SO_2}, X_{O_3}. All the models have a satisfactory result when the four regressors are considered. Such a conclusion is justified by the mean squared prediction error (MSPE) defined by

$$MSPE = \frac{1}{164}\sum_{j=1}^{164}\left(Y_j - \widehat{Y}_j\right)^2. \qquad \text{(Eq.6)}$$

The MSPE errors are summarized in *Table 1*.

***Table 1.** The MSPE of the three models*

Covariates	FNR model	FRR model	FRER
$X_{NO}, X_{NO_2}, X_{O_3}$ Case	**2.32**	3.12	2.98
X_{SO_2}, X_{O_3} Case	4.53	**3.09**	3.22
$X_{NO}, X_{NO_2}, X_{SO_2}, X_{O_3}$ Case	3.28	**3.19**	3.53

Discussion

The real time air quality in Marylebone road was analyzed by using the functional statistical models. A sample of 364 observations was used to analyze the impact of the other pollutant gases on the ozone concentration. The results of this statistical analysis are displayed in *Figures 3–5* and the prediction errors are summarized in *Table 1*. According to these prediction results, it appears clearly that the efficiency of this prediction issue is related to two important parameters which are the predictor model and the regressor variables. Precisely, using the three covariates $X_{NO}, X_{NO_2}, X_{O_3}$, Figure 3 shows that the best model is FNR; it gives an MSPE equal to 2.32. However, if we consider as covariates the two variables X_{SO_2}, X_{O_3} or $X_{NO}, X_{NO_2}, X_{SO_2}, X_{O_3}$ the best results were obtained by FRR model. Its prediction error is 3.09 and 3.19, respectively. Overall, it is clear that this functional regression models have satisfactorily performance. The prediction errors evaluated by the rule $MSPE$ is varied between 2.32-3.28 for the FNR, in the interval (3.12, 3.19) and is between 2.98-3.53 for the FRR and

FRER, respectively. All these predictor models are performed with fixed horizon equal to 4 h. Of course the choice of the prediction horizon is crucial. It should be enough for alarm purposes. Conceptually, our approach can be used for any horizon, but, we focused, here in 4 h ahead prediction, only, to show the easy implementation of the proposed algorithms. Furth more, the prediction results show that the three proposed functional models are fast, robust and accurate in this prediction issue. Moreover, the main feature of these functional models is the modelling of the daily curves of the pollutant gases in theirs continuous path allowing to explore the whole existing information of this time series data. Of course this consideration permits to avoid several drawbacks of the multivariate regression models, such as the curse of dimensionality or the calibration problem. On the other hand, the prediction in advance of the future values of the ozone concentration has a great importance for the decision-makers. They permit to plan the prevention strategy in order to combat the principal cause of the pollution. The present study contribute in this fundamental issue with these flexible models which can be used for various covariates variables and several pollutant factors, such as temperatures, wind (speed and directions), radiations, etc. Thus, the integration of these additional factors in our functional model is one of the natural prospects of this work. In addition, we can, also, predict the ozone concentration using other recent statistical models such as the partial functional linear model or the functional local linear models.

Conclusion

In this work, we have developed a new approach to predict the maximum ozone concentration using other air quality factors. These models are based on real-time measurement of the input data. The main feature of these models is the fact that they allow exploring all the information of these continuous-time observations viewed as curves. They are easily implementable, and their efficiency is related to the choice of the exogenous variables. In the sense that in the function of the considered covariates we can choose the adapted model. Moreover, the proposed models permit to avoid the core drawback of the classical models that is the loss of information after the prediction transformation. Indeed, the classical models induced by the multivariate regression are obtained by discretization of the daily curve in some finite grid or by projection into a finite-dimensional space. All these transformations on the input data are performed independently to the output variables. But, the ozone prediction is very sensitive to the input data. So the classical models are failed in these situations of real-time measured data. Thus, the originality of the functional approach comes from the fact that the prediction problem is performed without any transformation of the data. In conclusion, we can say that the accuracy of air quality prediction is based on both: the choice of the appropriate statistical models and the determination of the covariates. Moreover, the performance of the proposed models can be improved by using other covariates such as the metrological data wind direction, wind speed, humidity, solar radiation, air temperature.

Acknowledgments. Authors greatly thank the Editor in chief, the Associate Editor and the anonymous referees for a careful reading of the paper and for their valuable comments and suggestions which improved the quality of this paper substantially. They also extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through General Research Project under grant number (G.R.P-125-40).

REFERENCES

[1] Aneiros, G., Cao, R., Fraiman, R., Vieu, P. (2019): Editorial for the special issue on functional data analysis and related topics. – J. Multivariate Anal. 170: 1-3.

[2] Azzedine N., Laksaci A., Ould-Saïd, E. (2008): On robust nonparametric regression estimation for a functional regressor. – Statistics & Probability Letters 78: 3216-3221.

[3] Attouch Mk, Laksaci A., Said, O. (2012): Robust regression for functional time series data. – Journal of the Japan Statistical Society 42: 125-143.

[4] Bassoudi, M., Kaid, Z. (2018): Nonparametric relative error regression for functional ergodic data. – Int. J. Math. Stat. 19: 90-99.

[5] Demongeot J., Hamie A., Laksaci A., Rachdi, M. (2016): Relative-error prediction in nonparametric functional statistics: theory and practice. – Journal of Multivariate Analysis 146: 261-268.

[6] Ding W., Zhang J., Leung Y. (2016): Prediction of air pollutant concentration based on sparse response back-propagation training feedforward neural networks. – Environ Sci Pollut Res 23: 19481-19494.

[7] Ferraty, F. (2010): High-dimensional data: a fascinating statistical challenge. – Journal of Multivariate Analysis 101: 305-306.

[8] Ferraty, F., Vieu, P. (2006): Nonparametric Functional Data Analysis. Theory and Practice. – Springer Series in Statistics, New York.

[9] Ferraty F., Laksaci A., Tadj A., Vieu, P. (2010): Kernel regression with functional response. – Electronic Journal of Statistics 5(140): 335-352.

[10] Gavrila, C., Teodorescu, N., Gruia (2013): Bayesian modelling to the water loss management decisions. – Journal of Water Science and Technology: Water Supply 13: 883-888.

[11] Gavrila, C., Coman, A., Gruia, I., Ardelean, F., Vartires, A. (2016): Prediction method applied for the evaluation of the tropospheric ozone concentrations in Bucharest. – Romanian Journal of Physics 61(5-6): 1067-1078.

[12] Gong B., Ordieres-Meré, J. (2016): Prediction of daily maximum ozone threshold exceedances by preprocessing and ensemble artificial intelligence techniques: case study of Hong Kong. – Environ Model Softw 84: 290-303.

[13] Hsing, T., Eubank, R. (2015): Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators. – Wiley Series in Probability and Statistics. John Wiley & Sons, Chichester, UK.

[14] Ling, N., Vieu, P. (2018): Nonparametric modelling for functional data: selected survey and tracks for future. – Statistics 52: 934-949.

[15] Nghiem, L. H., Oanh, N. T. H. (2009): Comparative analysis of maximum daily ozone levels in urban areas predicted by different statistical models. – Science Asia 35: 276-283.

[16] Ryan, W. F. (1995): Forecasting severe ozone episodes in the Baltimore Metropolitan Area. – Atmospheric Environment 29: 2387-2398.

[17] Slini, T., Karatzas, K., Moussiopoulos, N. (2002): Statistical analysis of environmental data as the basis of forecasting: an air quality application. – The Science of the Total Environment 288: 227-237.

[18] Taylan, O. (2017): Modelling and analysis of ozone concentration by artificial intelligent techniques for estimating air quality. – Atmos Environ 150: 356-365.

[19] Yi, J., Prybutok, V. R. (1996): A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. – Environmental Pollution 92: 349-357.

[20] Zhang Y., Bocquet M., Mallet, V., Seigneur C., Baklanov, A. (2012): Real-time air quality forecasting. Part I: History, techniques, and current status. – Atmos Environ 60: 632-655.

DOI: http://dx.doi.org/10.15666/aeer/1706_1423114239