

# THE EFFECTS OF INDUSTRY INCREASE AND URBANIZATION ON AIR POLLUTANTS IN TURKEY: A NONLINEAR AIR QUALITY MODEL

KILINC, B. K.

*Department of Statistics, Faculty of Science, Eskisehir Technical University, Eskisehir 26470, Turkey*

*(e-mail: bkan@eskisehir.edu.tr; phone: +90-222-335-0580; fax: +90-222-320-4910)*

(Received 29<sup>th</sup> Mar 2019; accepted 24<sup>th</sup> May 2019)

**Abstract.** Daily concentrations of air pollutants greatly affect air quality as an increase in industry and urbanization deteriorate the environment. For forty cities in Turkey, eleven variables are recorded to investigate the determinants of air pollution presumably by regressing the air pollutants PM<sub>10</sub>, NO<sub>x</sub>, and NO<sub>2</sub>. The temperature, wind, human factors such as population, vehicles, manufacturer, and suchlike are used to create a nonlinear air quality model in Turkey due to the multivariate nature of the data. A comparison of the nonparametric models of the concentration of these pollutants, using multivariate adaptive regression splines (MARS), was obtained to estimate the dependence between air pollutants and various factors. Finally, a model for PM<sub>10</sub> concentration shows that climate effects are the most significant variables, whereas the predicted models for NO<sub>x</sub> and NO<sub>2</sub> indicate that human factors, such as the number of manufacturers and the number of vehicles, are significant variables. In conclusion, the predicted models are easy to interpret and have advantages of capacity to produce the contributions of the factors for each pollutant model. It is advisable for researchers to examine and determine the suitability of their data sets using nonlinear models when atypical observations and high correlations exist in the data.

**Keywords:** *multivariate adaptive regression splines, nonparametric, multicollinearity, outlier, pollutant*

## Introduction

The exposure of ambient air pollutants in industrial or metropolitan cities adversely affect humans, animals, plants, food crops or other living creatures (Lutgens and Tarbuck, 2001; Kelly and Fussell, 2015, 2017). The measurements of meteorological pollution, such as NO<sub>2</sub> (nitrogen dioxide), and PM<sub>10</sub> (particulate matter) have become more important due to their harmful effects on human health (Akkoyunlu, 2003; Elbir et al., 2000; Garcia, 2001; Godish, 2004; Garcia, 2006; Lutgens and Tarbuck, 2001). For example, the adverse effects on human health may include coughing, asthma, chronic bronchitis, and cardio vascular morbidity (Lutgens and Tarbuck, 2001; U.S. Department of Interior; Wark et al., 1997; Wang et al., 2004). Ifran and Shaw (2017) investigated the relationship between environmental pollution, energy consumption and the level of urbanization in South Asian countries using by a nonparametric additive model. The European Commission, World Health Organization (WHO) and various national/international environmental agencies have published their standards and air quality guidelines for allowable levels of air pollutants (Cooper and Alley, 2002; Lim, 2005; Surez et al., 2011; Wang et al., 2004).

According to WHO statements, 1.4 out of 2.4 million deaths each year are caused by indoor air pollution while the remaining deaths are due directly to attributable air pollution (Lutgens and Tarbuck, 2001; Wark et al., 1997; Wang et al., 2004). The WHO Air Quality Guidelines provide a Global Update of 205 for threshold values of air pollution. In addition, it reports that 91% of the world population live in places where WHO air quality levels were not met. The WHO guidelines are currently under revision

with an expected new publication in 2020. The latest Global conference on air pollution was organized in Geneva on 20 November 2018. Almost 900 participants registered for the conference, including 70 commitments from countries, cities and organizers discussing how to tackle climate change and related health considerations.

Furthermore, publications from the European Union (EU) regarding emissions of the main air pollutants monitored at more 2500 stations throughout Europe present updated data and its impact. Policies and air quality regimes are aimed at providing the lowest target concentrations of possible pollutants (Holnicki et al., 2017; Isikli et al., 2015). The guidelines apply worldwide and are based on expert evaluations for PM10 and NO<sub>2</sub> (see *Table 1*).

**Table 1.** Comparison of limit values

Average period	EU	Turkey (2017)	Turkey (2018)	Turkey (2019)	Permitted exceedances each year	
					EU (2019)	Turkey (2019)
PM10 (24 h)	50 µg/m <sup>3</sup>	70 µg/m <sup>3</sup>	60 µg/m <sup>3</sup>	50 µg/m <sup>3</sup>	35 times/year	35 times/year
NOX (1 year)	30 µg/m <sup>3</sup>	-	-	30 µg/m <sup>3</sup>	-	-
NO2 (1 h)	200 µg/m <sup>3</sup>	270 µg/m <sup>3</sup>	260 µg/m <sup>3</sup>	250 µg/m <sup>3</sup>	18 times/year	18 times/year

(Source: <http://ec.europa.eu/environment/air/quality/standards.htm>)

PM10, which is often a proxy indicator, for air quality refers to fine particles with a diameter of 10 microns or less, and are tiny solid particles and liquid suspended in a gas. It affects living organisms on a daily basis or over time. As the sources of PM10 can be humans, fossil fuel combustion (primarily from road transport), factories or natural, the increase of PM10 levels can be a cause of mortality or morbidity (Kelly and Fussell, 2015).

The potential risk of PM10 air pollution is critical in certain cities in Turkey, considering that the measured daily PM10 concentrations should be in the range of 50 µg/m<sup>3</sup> from January 1, 2019, due to EU Commission regulations (Istanbul Metropolitan Municipality).

Serious risks are posed to the habitat of humans, animals and plants, not only from exposure of PM10, but also from NO<sub>x</sub> (Shi et al., 2014). This is a general term that is used to describe certain gases, namely NO<sub>2</sub> and NO. NO<sub>2</sub> is especially emitted from high temperature combustion processes (power generation, engines in vehicles, heating, and suchlike) and are also produced naturally during thunderstorms by electric discharge. NO<sub>2</sub>, a reddish- brown toxic gas, has a sharp and biting odour. The initial product formed is NO (nitric oxide). When NO oxidizes further in the atmosphere, nitrogen dioxide (NO<sub>2</sub>) forms. Similarly, studies on NO<sub>2</sub> show that it is mainly due to the eutrophication of lakes and marine environments and acidification of terrestrial ecosystems (Greenfelt et al., 1994).

The concentration levels of NO<sub>2</sub> can also play a significant role in asthma and lung functions. Shao et al. (2006) show that urbanization and industrial developments cause deteriorating air quality. The results of their study are that emissions of NO<sub>2</sub> are from industrial and domestic energy production and transportation. When NO<sub>2</sub> is combined

with water, the main component of acid rain, which is one cause for deforestation, is observed (WHO).

With the negative effects of growing population trends and urbanization, the role of such atmospheric pollutants has received increasing attention. Many scientists study on global and regional estimations for such emissions. Seo et al. (2018) examined meteorological conditions together with emissions of air pollutants using long-term measurements in Seoul using by a multiple linear regression model. Sekulic and Kowalski (1992), Friedman and Roosen (1995) used three MARS models for NO<sub>2</sub>, SO<sub>2</sub>, and PM10 as a function of other measured relevant pollutants in air quality; nitric oxide (NO), carbon monoxide (CO), and ozone (O<sub>3</sub>). Donnelly et al. (2017) used a hybrid model including nonparametric regression to describe nonlinear variations in concentration levels with speed, direction and meteorological variables which are employed as inputs to a multiple linear regression. Tao et al. (2017) studied the impact of external pollution and human factors in the forest ecosystems.

It is often assumed that multivariate data should follow a known specific statistical distribution. However, the complex structure of the data and the relationship among various factors may be too complicated to explain general patterns or to model dependency between variables. When examining ecological data, the interaction of several environmental factors have to be recognized carefully and properly analysed.

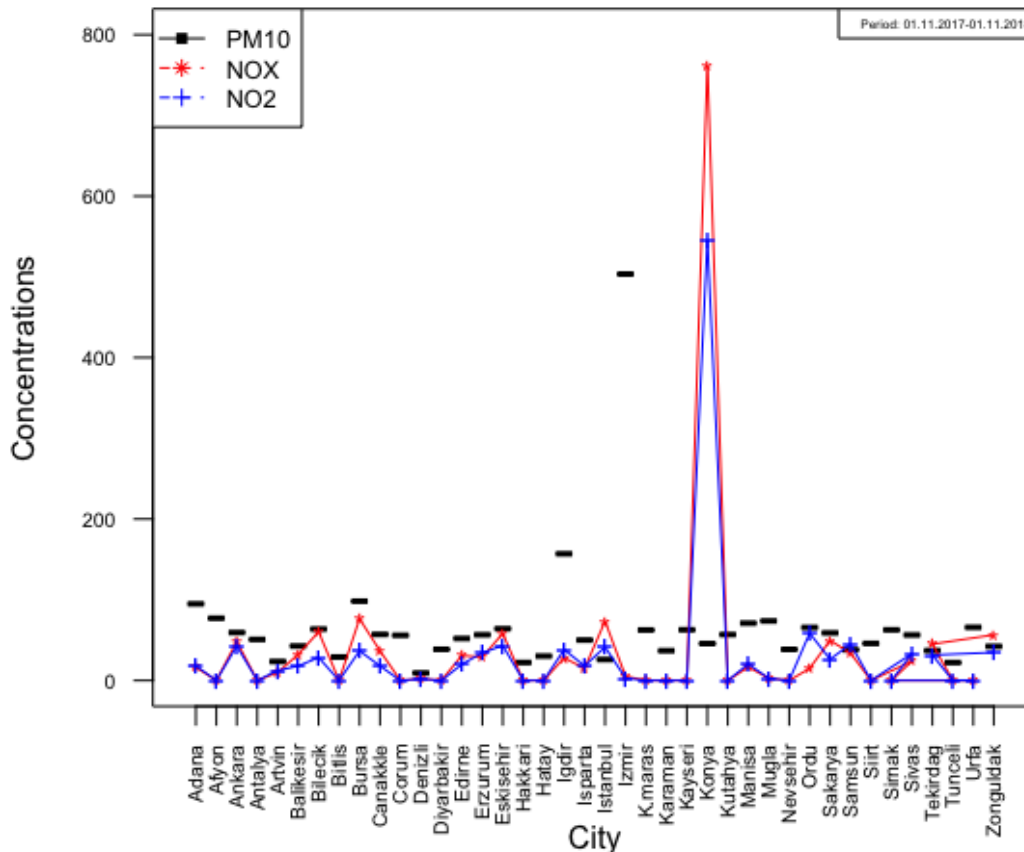
The aim of this study is to investigate the significant variables by using daily patterns of PM10, NO<sub>x</sub> and NO<sub>2</sub> and to provide meaningful future informative models for air quality control and regimes in Turkey. This study is organized as follows. Firstly, the current state of knowledge based on air pollutants in the literature is explained. Secondly, the motivations and data collection concerning our study are described. Next, a novel exploratory modelling technique, multivariate adaptive regression splines (MARS), is introduced by Friedman (1991). The results and comparisons from MARS models for each pollutant follow in the next section. Finally, the highlights of the paper are summarized in the conclusion.

### ***The motivation***

One major environmental risk is the concentration of air pollutants due to their harmful influence on health of humans, animals and plants or other living organisms. Low or middle income countries are exposed to pollutants in and around homes from the heating, lighting, traditional stoves used for cooking, polluting fuels on open fires and so on (WHO, 2018). The European Environmental Agency (EEA) indicate that there is an increasing trend of pollutants in urban regions. The WHO Air Quality 2005 document provides a guideline for an assessment of the health effects of air pollutants and guideline values for health risk pollution levels. The EU establishes the standards and objectives for a number of air pollutants in the air to be applied over different periods of time. Although Turkey and Switzerland are not the members of the EU, Turkey at least aims to adapt to EU standards by reaching the air quality limits. *Table 1* shows the limits aimed for in order to achieve the lowest possible concentrations of PM10, NO<sub>x</sub>, and NO<sub>2</sub>.

There are a number of examples that can be used as sources of air pollution including industrial activity, energy (heating, cooking, lighting, electricity), transportation (vehicles and fuels), waste reduction of forest pollution, increasing numbers of inhabitants, and climate. Addressing the risk factors for air pollution, countries could reduce levels of air pollution to safeguard the public health.

The Ministry of the Environment and Urbanisation is the main institution of the Turkish government that is responsible for transboundary atmospheric pollution control (Yurtseven et al., 2018). The measurements of PM<sub>10</sub>, NO<sub>x</sub> and NO<sub>2</sub> in forty cities organized from the Air Quality Monitoring System are shown in *Figure 1*. These are used to evaluate the priority of the variables that may be sources of the corresponding pollutants. Next, predictions from MARS models are depicted correctly as well as those cities with high levels of pollutants.



*Figure 1. Scatterplot of the data*

*Figure 1* shows concentrations of PM<sub>10</sub>, NO<sub>x</sub>, and NO<sub>2</sub> observed for the forty cities between 1 November, 2017, and 1 November, 2018. It is clear that the emissions peaks of PM<sub>10</sub>, NO<sub>x</sub> and NO<sub>2</sub> occur in some of the cities. Although the trend of the PM<sub>10</sub>, NO<sub>x</sub>, NO<sub>2</sub> is a quasi-sinoidal curve with peaks and valleys, these gases reach the greatest concentrations in Izmir and Konya, respectively. Additionally, certain relationships among variables are given in *Figure 2*.

## Materials and methods

### Model construction

The measurement of climatic variables is organized by the Ministry of the Environment and Urbanization Measurements and Monitoring Office, Turkey. For forty cities, eleven variables are recorded to investigate the determinants of air pollution

presumably by regressing PM10, NO<sub>x</sub>, NO<sub>2</sub> on those related to human ecology and climate. The measurements are collected at monitoring stations located within an industrial area close to the city or within the city limits. The results relating to climate (temperature, wind speed, precipitation, and number of days with precipitation) are average daily measurements (Air Quality Monitoring System).

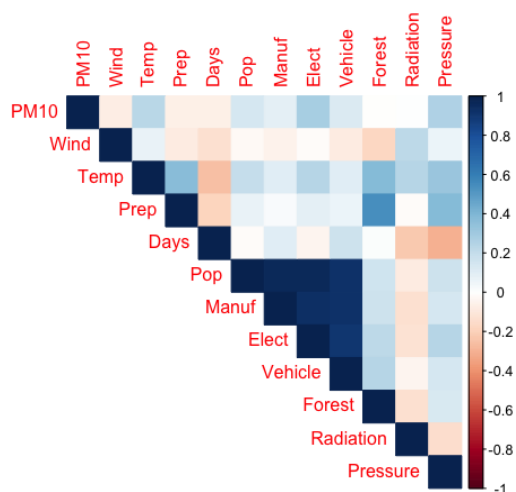


Figure 2. Correlations among variables

Addition, the effects of car pollution are widespread and is one of major cause of global warming. The increasing number of people which negatively effects the atmosphere and motorized vehicles can be geo-demographic predictors of the air pollutions and may be useful include in the model. The measurements for Radiation and Pressure are arranged from Turkish State Meteorological Service. The complete list of the data set was created using publicly online related web sources and the names of the variables are given in Table 2.

Table 2. Complete list of variables used in this study

Factors	Name of the variable	Mean	Min	Max
Average daily temperature period (°C)	Temp	13.6	5.7	19.1
Average daily wind period (km/h)	Wind	2.533	0	20.2
Average daily precipitation (kg/m <sup>2</sup> )	Prep	643.9	106.5	1222.1
Average number of days with precipitation	Days	110.1	73.9	438.4
Number of manufacturing enterprises consuming fuel oil	Manuf	218.9	2	2832
Population size (2017)	Pop	1487978	82498	15029231
Electricity consumption (%)	Elect	0.837	0.06	16.719
Number of motorized vehicles (2017)	Vehicle	270708	2048	2755250
Forest area - % of city area (2017)	Forest	32.8	0	68
Radiation in 2016 (kWh/m <sup>2</sup> )	Radiation	4.525	3.74	5.21
Average daily pressure	Press	755.4	0	1351.5
<b>Pollutants:</b>				
Particulate matter (µg/m <sup>3</sup> PM10)	PM10	65.14	9.26	503.47
NO <sub>x</sub> (µg/m <sup>3</sup> )	NO <sub>x</sub>	38.15	0	760.56
Nitrite concentration (µg/m <sup>3</sup> NO <sub>2</sub> )	NO <sub>2</sub>	29.3	0	544.81

### Multivariate adaptive regression splines

MARS is a nonparametric approximation of the relationship between a response and a set of independent variables in reflected pairs of simple linear splines (Friedman, 1991). The reflected pairs take the form given in *Equation 1*:

$$(x - t)_+ = \begin{cases} x - t & , x > t \\ 0 & \text{else} \end{cases} \quad (\text{Eq.1})$$

where the subscript '+' means the argument is a truncated power function.

MARS models do not assume any underlying relationship between the dependent variable and the predictors. This relationship is constructed by an adaptive fitting procedure where piecewise polynomials of degree  $q$  (splines) are driven from regressing the dependent variable onto predictors. The MARS model has the following form given in *Equation 2*:

$$y = \beta_0 + \sum_{m=1}^M \beta_m h_m(\mathbf{x}) \quad (\text{Eq.2})$$

where  $h_m(\mathbf{x})$  are basis functions and  $\beta_0, \dots, \beta_M$ , the unknown parameters for  $m = 1, \dots, M$ . Once the basis functions are investigated then  $\beta_0, \dots, \beta_M$  are estimated by the ordinary least squares method. The basis functions can be represented in *Equation 3* by:

$$h_m^q(\mathbf{x}) = \prod_{k=1}^{K_m} H[s_{k,m}(x_{v(k,m)} - t_{k,m})]_+^q \quad (\text{Eq.3})$$

with involving the truncated power functions with polynomials of lower order than  $q$ , where  $K_m$  is the number of variables (interaction order) in the  $m^{\text{th}}$  basis expansion and the number of splits that give raise to  $h_m(\mathbf{x})$ .  $x_{v(k,m)}$  is the  $v^{\text{th}}$  variable,  $v(k,m)$  label the predictor variables,  $1 \leq v(k,m) \leq n$ ,  $t_{k,m}$  is a knot on each of the corresponding variables and represent values on these variables. The quantities  $s_{k,m}$  in *Equation 2* take on values  $+/-1$  and indicate the left/right sense of the associated step functions.

The MARS algorithm adaptively selects the basis function set by two iterative approaches; forward and backward selection. It uses the residual squared error in iterations to compare the partition points. The criterion used to set the final model is a modified generalized cross validation (GCV) of the one first proposed by Craven and Wahba in 1979. The difference between the two criteria comes from a penalty term that reflects the complexity of the model in MARS.

The MARS approach allows a nonlinear relationship over different intervals of the vector of the explanatory variables for modeling  $Y$  (Sephton, 2001). The relationship between the explanatory variables and the response is fitted by basis functions that are basically splines. The main idea behind MARS is to explore the relationship by splitting the explanatory variables over its region into several intervals and to transform the original input variables over the intervals. MARS fits a spline based model in each interval. Basis Functions ( $B_i$ ) include knot locations relating to the explanatory variables.  $B_i$  can be a single or multivariable interaction term. The final model is a combination of  $B_i$ . Model selection is accomplished using the GCV criterion given in *Equation 4*:

$$\text{GCV} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{f}_M(\mathbf{x}))^2 / \left(1 - \frac{C(M)}{N}\right)^2 \quad (\text{Eq.4})$$

where  $C(M)$  is the complexity parameter to the corresponding model, and  $N$  represents the total number of observations. The final MARS model is the one with the smallest GCV and the largest  $R_{adj}^2$ .

## Analysis and results

The conventional methods, multiple linear regression (MLR) and principal component analysis are used to address the main determinants of pollutants of PM10, NO<sub>x</sub>, NO<sub>2</sub>, respectively. Multiple linear regression is used by regressing several predictors listed in *Table 2* on the concentrations of PM10, NO<sub>x</sub>, and NO<sub>2</sub>. The fitted values did not produce statistically significant models except the PM10-model ( $F_{PM10;11;28} = 3.002$ ,  $p_{PM10} = 0.0092$ ,  $R^2\text{-adj} = 0.36$ ;  $F_{NO_x} = 1.565$ ,  $p_{NO_x} = 0.164$ ,  $R^2\text{-adj} = 0.13$ ;  $F_{NO_2} = 1.388$ ,  $p_{NO_2} = 0.232$ ,  $R^2\text{-adj} = 0.23$ ). It is also clear from the multiple plots given in the *Appendix* that MLR can not be considered for further analysis.

Next, principal component regression (PCR) is performed to investigate the determinants of the pollutants. The models from PCR did not provide informative contribution in this analysis and are effected by the presence of the outliers. Although the first four components (four eigenvalues greater than 1) explained 84.3% of the total variation among the transformed predictors, the regression models were not statistically found significant ( $F_{PM10;4;35} = 0.721$ ,  $p_{PM10} = 0.583$ ;  $F_{NO_x;4;35} = 0.951$ ,  $p_{NO_x} = 0.446$ ;  $F_{NO_2;4;35} = 1.002$ ,  $p_{NO_2} = 0.419$ ). Ultimately, PCR was not appropriate for predicting the responses on several variables.

In this work, a new approach for modelling PM10, NO<sub>x</sub>, and NO<sub>2</sub> concentrations are developed. MARS models have been used (Milborrow, 2011; Hastie et al., 2003), the basis functions of the models, consisting of linear and second-order splines, are shown in *Tables 3, 4, and 5*. There are three dependent variables used to build the MARS models; particulate matter less than 10  $\mu\text{g}/\text{m}^3$  (PM10), nitrogen dioxide (NO<sub>2</sub>), and nitric oxide (NO<sub>x</sub>). The results of MARS computed using all the available observations show a list of 4, 9, and 6 basis functions ( $B_i$ ) for each of the three MARS models and their coefficients ( $C_i$ ), respectively.

It should be noted that, MARS constructs nonparametric regression models as an extension of the linear models, and automatically determines the linearities and interactions and models a weighted sum of basis functions (hinge functions). The hinge function  $h()$  is  $x$  if the first quadrant is greater than 0, and is 0 if  $x$  is less than or equal to 0.

The regression equation of PM10, which includes linear form of its term, such as Precipitation and two second order basis functions, can be easily generated using *Table 3* and *Equation 5* as follows:

$$PM10 = 46.002 + 0.466 B_2 - 0.016 B_3 + 0.158 B_4 \quad (\text{Eq.5})$$

In *Equation 5*, there are two interaction terms,  $B_3$  and  $B_4$ , where certain relationships between the variables Pressure and Forest, and the variables Electricity and Pressure found to be important, respectively. Using *Equation 5*, PM10 predictions can be calculated easily. The variation between the predictions and the actual data is presented in *Figure 4*. Similarly, the regression equations for NO<sub>x</sub> and NO<sub>2</sub> can be constructed easily from *Tables 4* and *5*.

**Table 3.** List of basis function and their coefficients obtained by the MARS model for PM10 pollutant

<b>C<sub>i</sub></b>	<b>Definition</b>	<b>B<sub>i</sub></b>
46.002	1	B <sub>1</sub>
0.466	h(452.3-Prep)	B <sub>2</sub>
-0.016	h(452.3-Prep)*Forest	B <sub>3</sub>
0.158	Elect* h(Pressure-988.77)	B <sub>4</sub>

It is worth to notice that the MARS model for NO<sub>x</sub> allows us to assess many interactions between different hinge functions. This can be easily seen in *Table 4*. The NO<sub>x</sub> model includes three interaction terms between hinge functions and linear terms. For example,  $h(13.8-Temp)*Vehicle$  is an interaction effect for those Temp values less than 13.8 and the number of vehicles. Also, the MARS model for NO<sub>x</sub> includes linear form of its term such as Vehicle and hinge functions of the variables, Vehicle, Manuf, and Pressure.

**Table 4.** List of basis function and their coefficients obtained by the MARS model for NO<sub>x</sub> pollutant

<b>C<sub>i</sub></b>	<b>Definition</b>	<b>B<sub>i</sub></b>
4.07654	1	B <sub>1</sub>
0.00077	Vehicle	B <sub>2</sub>
-11.01813	h(Manuf-151)	B <sub>3</sub>
-0.00064	h(Vehicle-86926)	B <sub>4</sub>
-0.59119	h(Pressure-988.77)	B <sub>5</sub>
-0.00016	h(13.8-Temp)*Vehicle	B <sub>6</sub>
-0.00013	h(Temp-13.8)*Vehicle	B <sub>7</sub>
2.63176	h(Manuf-151)*Radiation	B <sub>8</sub>
-0.00026	h(988.77-Pressure)	B <sub>9</sub>

According to the results of NO<sub>2</sub>, the model is constructed by linear and second-order basis functions. There are two linear basis functions and three second-order basis functions where certain relationships between Prep and the variables Manuf, Vehicle, and Pressure are found to be statistically important.

**Table 5.** List of basis function and their coefficients obtained by the MARS model for NO<sub>2</sub> pollutant

<b>C<sub>i</sub></b>	<b>Definition</b>	<b>B<sub>i</sub></b>
2.75336	1	B <sub>1</sub>
-0.32477	h(581.8-Prep)	B <sub>2</sub>
47.19741	h(4.66-Radiation)	B <sub>3</sub>
0.00441	h(581.8-Prep)*Manuf	B <sub>4</sub>
-0.000002	h(581.8-Prep)*Vehicle	B <sub>5</sub>
0.000304	h(581.8-Prep)* Pressure	B <sub>6</sub>



The most significant variables in the prediction of PM<sub>10</sub>, NO<sub>x</sub> and NO<sub>2</sub> can be obtained from each MARS model as well. The list of each model is shown in *Tables 6, 7, and 8*. These tables list the values of three criteria for estimating variable importance. The *nSubset* criterion indicates the counts of those models that include the variable. More subsets are considered of greater importance. The residual sum of the squares (RSS) criterion calculates the decrease in the RSS for each subset. The variables which cause a large decrease in RSS are considered more important. The last criterion GCV is similar.

**Table 6.** Evaluation of the variable importance for the PM<sub>10</sub> pollutant model according to *Nsubsets, GCV and RSS*

Variable	nSubsets	GCV	RSS
Elect	3	100	100
Pressure	3	100	100
Prep	2	15.5	21.9
Forest	1	14.1	16.66

RSS: residual sum of squares; GCV: generalized cross validation; Nsubsets: number of subset models in which variable occurs

The most important variables in the PM<sub>10</sub> prediction given in *Table 6* are obtained from electricity, pressure, precipitation, and forest in order of priority.

**Table 7.** Evaluation of the variable importance for the NO<sub>x</sub> pollutant model according to *Nsubsets, GCV and RSS*

Variable	nSubsets	GCV	RSS
Manuf	8	100	100
Radiation	8	100	100
Temp	6	74.8	67.6
Vehicle	6	74.8	67.6
Pressure	3	28.5	26.3

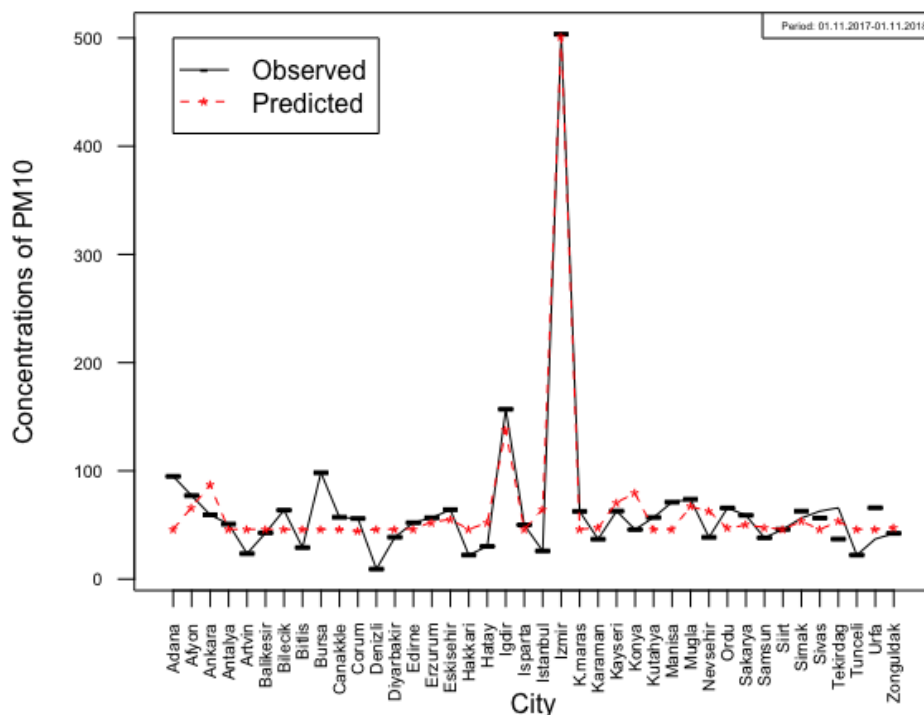
The results in *Table 7* indicate that the most significant variables are Manufacturer, Radiation, Temperature, Vehicle and Pressure in the prediction of NO<sub>x</sub>.

In *Table 8*, the most significant variables in the NO<sub>2</sub> prediction are Precipitation, Manufacturer, Vehicle, Pressure and Radiation.

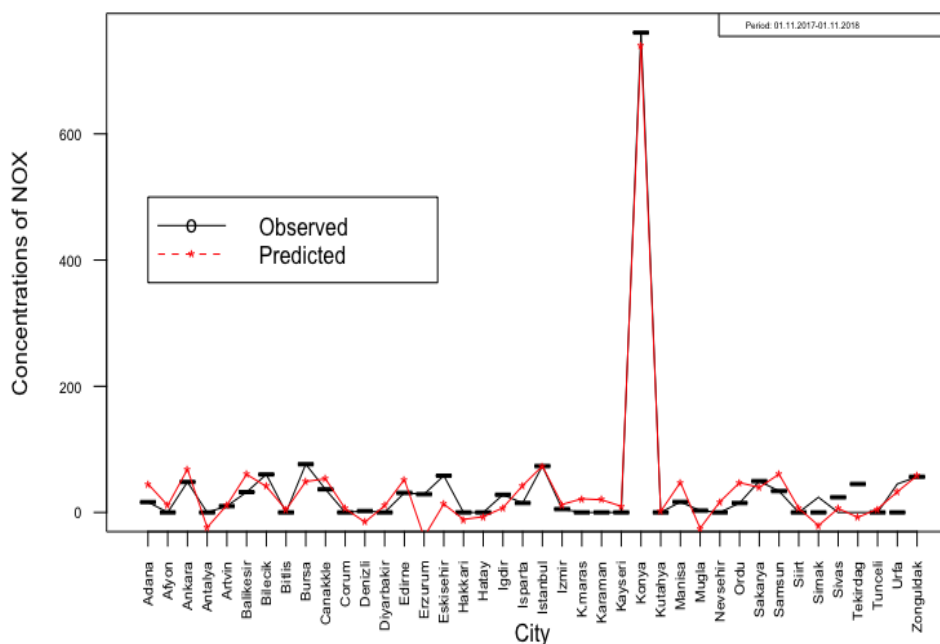
**Table 8.** Evaluation of the variable importance for the NO<sub>2</sub> pollutant model according to *Nsubsets, GCV and RSS*

Variable	nSubsets	GCV	RSS
Prep	5	100	100
Manuf	5	100	100
Vehicle	5	87.9	82.6
Pressure	1	23.2	22.0
Radiation	1	15.8	15.8

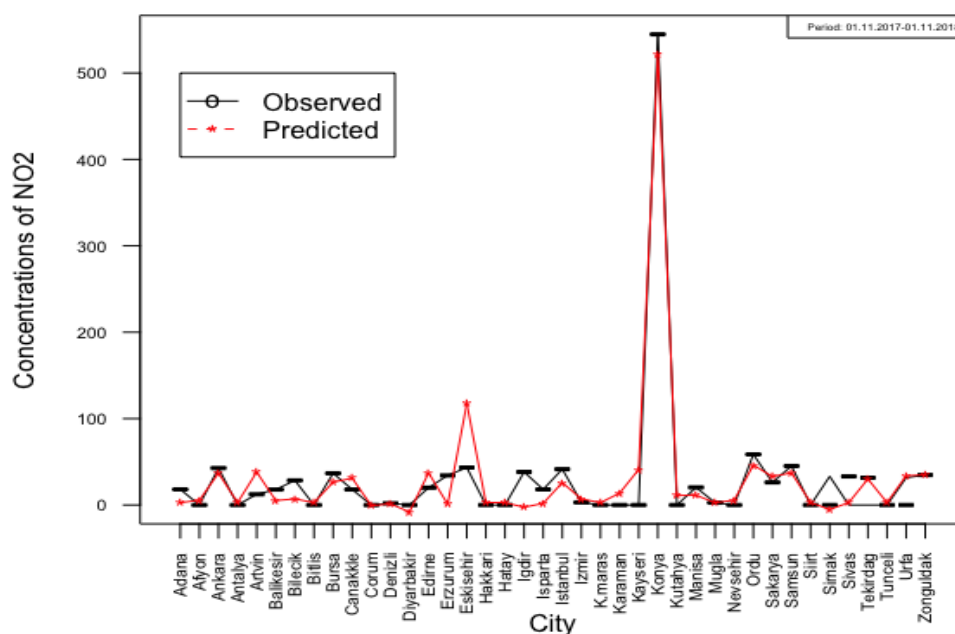
Furthermore, comparisons between concentrations of PM<sub>10</sub>, NO<sub>x</sub> and NO<sub>2</sub>, predicted by the MARS model and observed, are represented by *Figures 3, 4 and 5*, respectively.



**Figure 3.** Comparisons between concentrations of PM<sub>10</sub> predicted by the MARS model and those observed



**Figure 4.** Comparisons between concentrations of NO<sub>x</sub> predicted by the MARS model and those observed



**Figure 5.** Comparisons between concentrations of  $\text{NO}_2$  predicted by the MARS model and those observed

In Table 9, an evaluation of the GCV and  $R^2$  (coefficient of determination) of the models is summarized. A coefficient of determination value indicates how well a model explains and predicts the future. A value of 1.0 is assessed as a perfect fit. For the  $\text{PM}_{10}$ ,  $\text{NO}_x$  and  $\text{NO}_2$  models, the coefficients of determination are 92.5%, 96.2%, and 94.2%, respectively.

**Table 9.** Evaluation of the GCV and  $R^2$  that form the models

Variable	GCV	$R^2$
$\text{PM}_{10}$ -second order	665.114	0.925
$\text{NO}_x$ -second order	2294.915	0.962
$\text{NO}_2$ -second order	932.613	0.942

## Discussion and conclusion

In this research work, three models based on the multivariate adaptive regression splines for the study of considered pollutants  $\text{PM}_{10}$ ,  $\text{NO}_x$ , and  $\text{NO}_2$  are proposed. The main purpose of this study is to increase accuracy in the prediction of three air pollutants when human factors and climate effects are present and physical-chemical parameters are combined. The innovative methodology MARS is applied to a real data set and is able to predict successfully.

One conclusion of this study is to set the order of the most significant variables in the prediction of  $\text{PM}_{10}$ ,  $\text{NO}_x$ , and  $\text{NO}_2$ . The priority of the predictor variables involved in the estimation of  $\text{PM}_{10}$  are actually related to climate parameters. On the other hand, human factors such as Manufacturer and Vehicles, included in the prediction of  $\text{NO}_x$  and  $\text{NO}_2$ , are the most influential parameters followed by Precipitation and Forest, respectively. Furthermore, this paper presents simple statistical models for the

prediction of air pollutants. The results of numerical experiments, based on the application of the MARS technique, confirm high accuracy of daily modelling for pollutants PM<sub>10</sub>, NO<sub>x</sub>, and NO<sub>2</sub>.

Finally, the results of this study regarding the development of nonlinear models of three pollutant concentrations are valuable for future projects based on human health. Additionally, the nonlinear modelling of the ecological data with atypical observations can be used to reduce the effects of pollutants and to improve the development of models regarding air quality.

## REFERENCES

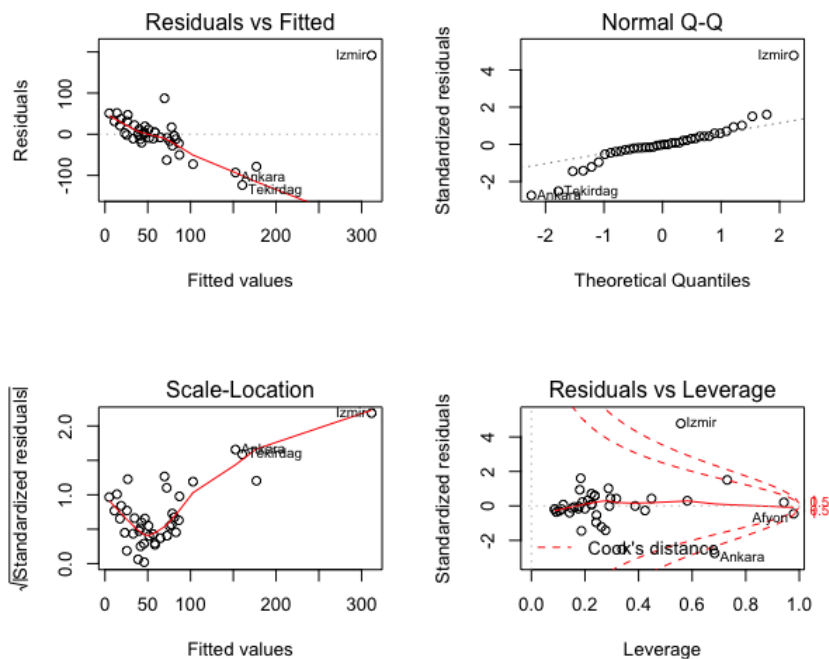
- [1] Air Quality Monitoring System: <http://www.havaizleme.gov.tr>. – Last Access: 01 November 2018.
- [2] Akkoyunlu, A., Ertürk, F. A. (2003): Evaluation of air pollution trends in Istanbul. – *International Journal of Environment and Pollution* 18: 388-398.
- [3] Craven, P., Wahba, G. (1979): Smoothing noisy data with spline functions. – *Numerische Mathematik* 31: 377-403.
- [4] Cooper, C. D., Alley, F. C. (2002): *Air Pollution Control*. – Waveland Press, New York.
- [5] Donnelly, A., Naughton, O., Broderick, B., Misstear, B. (2019): Short-term forecasting of Nitrogen dioxide (NO<sub>2</sub>) levels using a hybrid statistical and air mass history modelling approach. – *Environmental Modelling & Assessment* 22(3): 231-241.
- [6] Elbir, T., Muezzinoglu, A., Bayram, A. (2000): Evaluation of some air pollution indicators in Turkey. – *Environment International* 26(1-2): 5-10.
- [7] European Commission, Environment: <http://ec.europa.eu/environment/air/quality/standards.htm>. – Last Access: 11 December 2018.
- [8] European Environmental Agency: [https://www.eea.europa.eu/publications/corporate\\_document\\_2007\\_2](https://www.eea.europa.eu/publications/corporate_document_2007_2). – Last access: 05 January 2019.
- [9] Friedman, J. H. (1991): Multivariate adaptive regression splines. – *Annals of Statistics* 19: 1-141.
- [10] Friedman, J. H., Roosen, C. B. (1995): An introduction to multivariate adaptive regression splines. – *Statistical Methods in Medical Research* 4: 197-217.
- [11] García Nieto, P. J. (2001): Parametric study of selective removal of atmospheric aerosol by coagulation, condensation and gravitational settling. – *International Journal of Environmental Health Research* 11: 151-162.
- [12] García Nieto, P. J. (2006): Study of the evolution of aerosol emissions from coal-fired power plants due to coagulation, condensation, and gravitational settling and health impact. – *Journal of Environmental Management* 79(4): 372-382.
- [13] Grennfelt, P., Hov, O., Derwent, D. (1994): Second generation abatement strategies for NO<sub>x</sub>, NH<sub>3</sub>, SO<sub>2</sub> and VOC. – *Ambio* 23: 425-433.
- [14] Godish, T. (2004): *Air Quality*. – Lewis Publishers, Boca Raton, FL.
- [15] Hastie, T., Tibshirani, R., Friedman, J. H. (2003): *The Elements of Statistical Learning*. – Springer-Verlag, New York.
- [16] Holnick, P., Kaluszko, A., Nahorski, Z., Stankiewicz, K., Trapp, W. (2017): Air quality modelling for Warsaw agglomeration. – *Archives of Environmental Protection* 43(1): 48-64.
- [17] Irfan, M., Shaw, K. (2017): Modeling the effects of energy consumption and urbanization on environmental pollution in South Asian countries: a nonparametric panel approach. – *Quality & Quantity* 51: 65-78.
- [18] Isikli, E., Ustundag, A., Cevikcan, E. (2015): The effects of environmental risk factors on city life cycle: a link analysis. – *Human and Ecological Risk Assessment: An International Journal* 21(5): 1379-1394.

- [19] Istanbul Metropolitan Municipality: [www.ibb.gov.tr](http://www.ibb.gov.tr). – Last Access: 30 January 2019.
- [20] Lim, L. L., Hughes, S. J., Hellawell, E. E. (2005): Integrated decision support system for urban air quality assessment. – *Environmental Modelling & Software* 20: 947-954.
- [21] Lutgens, F. K., Tarbuck, E. J. (2001): *The Atmosphere: An Introduction to Meteorology*. – Prentice Hall, New York.
- [22] Kelly, F. J., Fussell, J. C. (2015): Air pollution and public health: emerging hazards and improved understanding of risk. – *Environmental Geochemistry and Health* 37(4): 631-649.
- [23] Kelly, F. J., Fussell, J. C. (2017): Role of oxidative stress in cardiovascular disease outcomes following exposure to ambient air pollution. – *Free Radical Biology and Medicine* 110: 345-367.
- [24] Milborrow, S. (2011): *Earth: Multivariate Adaptive Regression Splines*, R package. – Derived from *mda:mars* by T. Hastie and R. Tibshirani.
- [25] Ministry of Environment and Urbanization, Measurements and Monitoring Office: <https://www.mgm.gov.tr/kurumsal/istasyonlarimiz.aspx>. – Last Access: 01 November 2018.
- [26] Related Web Sources: <https://www.tobb.org.tr>, [www.enerjiatlas.com](http://www.enerjiatlas.com), [www.cmo.org.tr](http://www.cmo.org.tr), <http://otomobil.haber7.com>, <https://www.ogm.gov.tr>, <https://www.nufusu.com>.
- [27] Sekulic, S. S., Kowalski, B. R. (1992): MARS: a tutorial. – *Journal of Chemometrics* 6: 199-216.
- [28] Seo, J., Park, D. S. R., Kim, J. Y., Youn, D., Lim, Y. B., Kim, Y. (2018): Effects of meteorology and emissions on urban air quality: a quantitative statistical approach to long-term records (1999-2016) in Seoul, South Korea. – *Atmospheric Chemistry Physics* 18(21): 16121-16137.
- [29] Sephton, P. (2001): Forecasting recessions: can we do better on MARS? – *Review Federal Reserve Bank of St. Louis* 2001(March): 39-49.
- [30] Shao, M., Tang, X. Y., Zhang, Y. H., Li, W. J. (2006): City clusters in China: air and surface water pollution. – *Frontiers in Ecology and the Environment* 4: 353-361.
- [31] Shi, P., Xie, P. H., Qin, M., Si, F. Q., Dou, K., Du, K. (2014): Cluster analysis for daily patterns of SO and NO measured by the DOAS system in Xiamen. – *Aerosol and Air Quality Research* 14: 1455-1465.
- [32] Suárez Sánchez, A., García Nieto, P. J., Riesgo Fernández, P., del Coz Díaz, J. J., Iglesias-Rodríguez, F. J. (2011): Application of a SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain). – *Mathematical and Computer Modelling* 54(5-6): 1453-1466.
- [33] U.S. Department of the Interior, National Park Service, Last Access: 11 December 2018.
- [34] Tao, N., Liu, D., Wu, J. (2017): Evaluation study on ecosystem based on system dynamics: A case study of Gongga Mountain. – *Journal of Interdisciplinary Mathematics* 20(6-7): 1415-1418.
- [35] Yurtseven, E., Vehid, S., Bosat, M., Köksal, S., Yurtseven, C. M. (2018): Assessment of ambient air pollution in Istanbul during 2003-2013. – *Iranian Journal of Public Health* 47(8): 1137-1144.
- [36] Wang, L. K., Pereira, N. C., Hung, Y. T. (2004): *Air Pollution Control Engineering*. – Humana Press, New York.
- [37] Wark, K., Warner, C. F., Davis, W.T. (1997): *Air Pollution: Its Origin and Control*. 3<sup>rd</sup> Ed. – Prentice Hall, New Jersey.
- [38] World Health Organization (WHO) (2018): <https://www.who.int/en>. – Last Access: 11 December 2018.

## APPENDIX

Because of the nature of this multivariate data set, atypical observations are not easy to identify using univariate approaches. Therefore, we provide four plots; a Scale-Location plot of against fitted values, a Normal Q-Q plot, and a plot of Cook's distances versus row labels for each air pollutants model in *Figures A1, A2 and A3*.

*Figure A1. Multiple plots for PM10*



*Figure A2. Multiple plots for NOX*

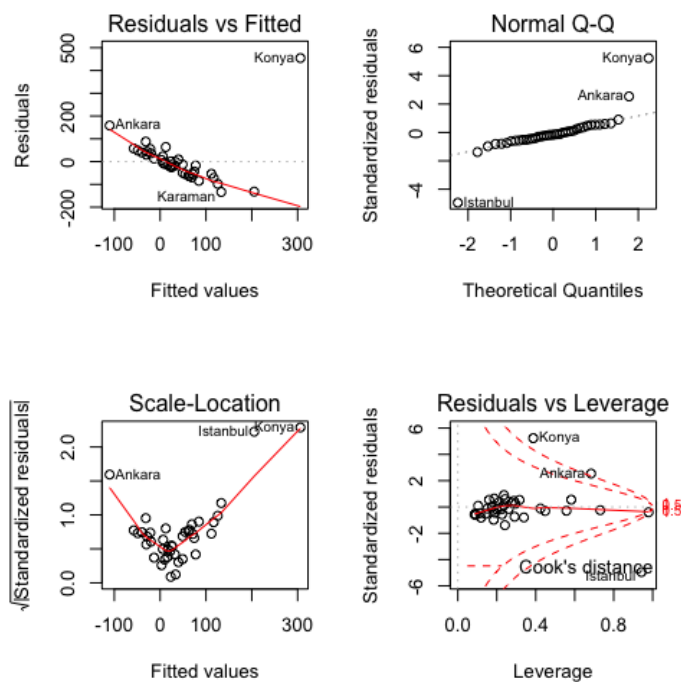
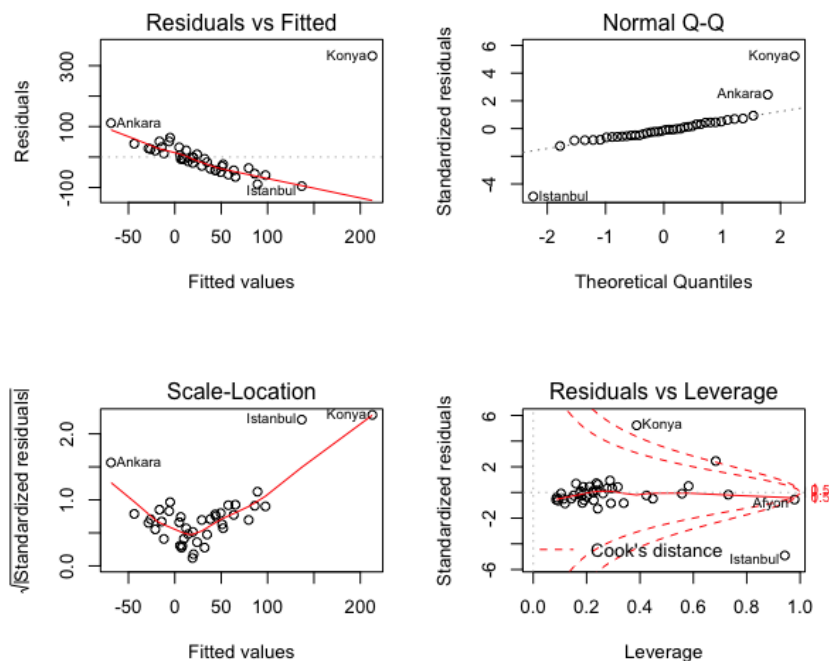


Figure A3. Multiple plots for NO<sub>2</sub>



In addition, VIF (Variance Inflation Factor) values for a possible linear model for each pollutant indicates higher degree of multicollinearity 53.4, 21.6, 15.3 and 27.8 for the variables Population, Electricity, Vehicle and Manufacturer, respectively.