Air Quality, Atmosphere & Health (2018) 11:1155–1165
https://doi.org/10.1007/s11869-018-0615-z

Comparing different methods for statistical modeling of particulate matter in Tehran, Iran

Vahid Mehdipour[1] · David S. Stevenson[2] · Mahsa Memarianfard[1] · Parveen Sihag[3]

Received: 18 May 2018 / Accepted: 8 August 2018 / Published online: 24 August 2018

Abstract

Particulate matter has major impacts on human health in urban regions, and Tehran is one of the most polluted metropolitan cities in the world, struggling to control this pollutant more than any other contaminant. $PM_{2.5}$ concentrations were predicted by three statistical modeling methods: (i) decision tree (DT), (ii) Bayesian network (BN), and (iii) support vector machine (SVM). Collected data for three consecutive years (January 2013 to January 2016) were used to develop the models. Data from the initial 2 years were employed as the training data, and measurements from the last year were used for testing the models. Twelve parameters, covering meteorological variables and concentrations of several chemical species, were explored as potential predictors of $PM_{2.5}$. According to the sensitivity analysis of $PM_{2.5}$ by SVM and derived explicit equations from BN and DT, PM_{10}, NO_2, SO_2, and O_3 are the most important predictors. Furthermore, the impacts of the predictors on the $PM_{2.5}$ were assessed which the chemical precursors' influences indicated more in comparison with meteorological parameters. Capabilities of the models were compared to each other and the support vector machine was found to be the best performing, based on evaluation criteria. Nonetheless, the decision tree and Bayesian network methods also provided acceptable results. We suggest more studies using the SVM and other methods as hybrids would lead to improved models.

Keywords Air pollution · Bayesian network · Decision tree · Support vector machine · Particulate matter

Introduction

Regarding a worldwide study accompanied by World Health Organization, it has been revealed that annually, three million people lose their lives due to severe air pollution (WHO 2003). Health scientists around the world have scrutinized air pollutant impacts on humans (Liu and Peng 2018; Pope et al. 2018; Li et al. 2018) and other living organisms and found that particulate matter with a diameter less than 2.5 μm ($PM_{2.5}$) is one the most detrimental pollutants (Davidson et al. 2005). $PM_{2.5}$ has been found to be one the most hazardous pollutants for human health in several studies (Sfetsos and Vlachogiannis 2010; Xing et al. 2016; Schweitzer and Zhou 2010; Cao et al. 2013; Borja-Aburto et al. 1998); hence, more attention and specific researches about $PM_{2.5}$ are required. Atkinson et al. (2014) studied a comprehensive, systematic review and meta-analysis of 110 published papers in health databases about $PM_{2.5}$, resulted that a 10 μg/m^3 increase of particulate matter concentration in an industrial city can cause and increase up to 2% in mortality due to cardiovascular and respiratory diseases. In a similar study on the nine Californian counties, the particulate matters' (PM_{10} and $PM_{2.5}$) impacts on the different parts of the society, with respect to sex, age, ethnicity, and so on of the members, have been analyzed. Results showed that 10 μg/m^3 increment in $PM_{2.5}$ concentration, only in 2 days, is the main factor of the 0.6% of the more mortality. These and other peer-reviewed studies about the particulate matter and human health (Pascal et al. 2013; Marzouni et al. 2016; Fattore et al. 2011; Fann et al. 2012; Dunea et al. 2016; Leili et al. 2008) illustrate the importance of more and accurate studies about $PM_{2.5}$. Thus, in this paper, we aimed to prognosticate the $PM_{2.5}$ concentration in Tehran, Iran, exploiting statistical modeling techniques. We used Bayesian network (BN) and decision tree (DT) as two of the reliable methods along with support vector machine (SVM) as a machine learning approach to predict the $PM_{2.5}$

✉ Vahid Mehdipour
vahid.mehdipour1992@gmail.com

[1] Department of Civil and Environment Engineering, K N Toosi University of Technology, Tehran, Iran

[2] School of GeoSciences, The University of Edinburgh, Edinburgh, UK

[3] National Institute of Technology, Kurukshetra, India

concentration and compared these three methods' capabilities to each other with respect to the statistical results. Exploiting "intelligent machines" for data mining and variable prediction is prevalent in all scientific topics, and for environmental parameters, these methods have given promising results (Martí et al. 2013; Sharifi et al. 2016; Mehdipour et al. 2017; Kim et al. 2015). Mehdipour (2017) compared four prominent methods: gene expression programming, support vector machine, artificial neural network, and wavelet to forecast ground level ozone (O_3) in Tehran. The results indicated that SVM has the best accuracy. Feng et al. (2015) studied the $PM_{2.5}$ prediction in the Beijing, Tianjin, and Hebei provinces in China during a year and have used the wavelet transformation and geographic model to improve the artificial neural network (ANN) accuracy, and they recommended their method to be implemented on other countries' air pollution centers. Wang et al. (2015) evolved a novel model for prediction of PM_{10} and SO_2 daily concentration. They used a Taylor expansion forecasting model to ameliorate the support vector machine and artificial neural network, and finally assessed their own new model as a very promising one. Kisi et al. (2017) applied least square support vector regression (LSSVR) and multivariate adaptive regression splines (MARS) and M5 Model Tree (M5-Tree) to forecast sulfur dioxide (SO_2) in three regions in India and the LSSVR had the best results. Decision tree and Bayesian belief networks have been applied to several environmental topics (McCann et al. 2006; Marchant and Ramos 2012; Liu et al. 2012; Aguilera et al. 2011) and their abilities for $PM_{2.5}$ prediction have been analyzed and compared with others. Kujaroentavon et al. (2015) introduced the decision tree to classify the air pollution in Thailand. They used the air quality index (AQI) and decision tree to classify the air pollution levels for human health and the results were satisfactory. McMillan et al. (2007) aimed to find a way to validate the air pollution data monitoring. The model specified in a Bayesian framework and fitted by Markov Chain Monto Carlo techniques. Vafa-arani et al. (2014) conducted a research by dynamic modeling for analyzing the most important factors on the Tehran air pollution. The technology improvement in fuel and automotive industry and public transportation are the most affective factors among other manifold choices such as industry-related parameters, road construction, traffic control plan, and urban transportation. In another study in Tehran, the ground level ozone (O_3) concentration has been scrutinized by Mehdipour and Memarianfard (2017) exploiting support vector machine and gene expression programming which are two most potent machine learning methods and the results comparing the predicted dataset with the testing ones, depicted acceptable upshots. All above cited papers show a deep indispensability for an accurate study about the fine particulate matter, and also support vector machine, decision tree, and Bayesian network have been introduced as potent methods for environmental problems.

Methodology

Decision tree

Decision tree (DT) is an expedient way to illustrate a concept which also is a tool for decision-making and we can evolve models to prognosticate a target value with respect to the input parameters and datasets (Rivest 1987). DT is a proper and prevalent method for data mining. We exploited a model or graph liken to tree showing an algorithm to find the best strategy with the most possibility to reach the target (Utgoff 1989). In decision analysis, a decision tree, specifically the diagram of the decision, represents a visible tool for more understandable and analytical decision-making (Kamiński et al. 2017). This tool classifies the "test" datasets from root up to branches and leaves. Every leaf of the tree represents a particular class. A well-developed tree is capable of handling of manifold parameters with numerous data for each parameter (Quinlan 2006). Three kinds of nodes are available in a graph of DT (Moret 1982): (a) decision node: square, (b) chance node: circle, and (c) end note: triangle. Every inner node corresponds an input data and the edges to children for each of the probable values of that input variable. A leaf depicts a value of the target variable given the values of the input variables represented by the path from the root to the leaf (James et al. 2000).

In this study, we developed a tree in which 12 predictors are assumed as the input values and $PM_{2.5}$ plays the target parameter's role. Wind speed, maximum ambient temperature, minimum ambient air temperature, average nebulosity, sunshine, humidity, participation, carbone monoxide, ground level ozone, nitrogen dioxide, sulfur dioxide, and particulate matter with 10-μm diameter used as predictors and the particulate matter with 2.5-μm size is the target variable. An expanded and wide tree may encounter with deep overfitting problem and a limited one probably cannot consider the all variables, where pruning the tree is a tool to keep the tree size in acceptable and optimum range. Overfitting occurs when the machine instead of learning memorizes the data sets and produces very similar outcomes to inputs.

Support vector machine

For the first time, Cortes and Vapnik (1995) invented a machine using vectors to classify the datasets into a two-dimensional space. Machines which use a part of datasets for training and another part for testing commonly categorize the datasets. According to Fig. 1, a vector machine can easily classify the datasets into groups in two-dimensional ambient by myriad cross lines and a particular super line. The best separating line or super line has the maximum distance from the border lines. Equations 1 and 2 represent the borderlines,

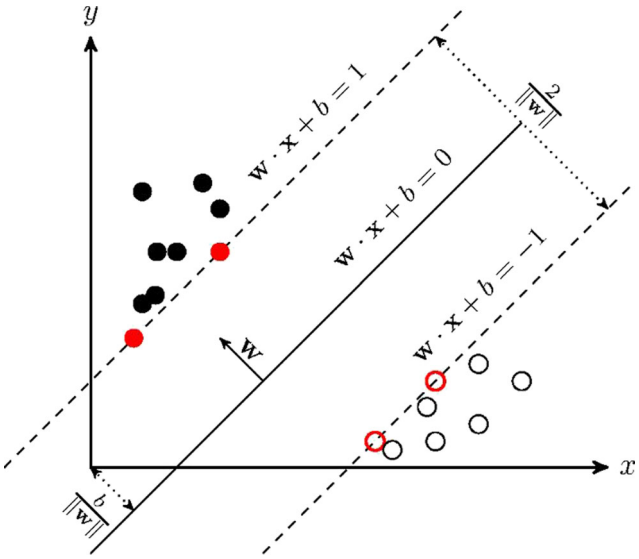

Fig. 1 Boarder lines (support vectors) and the super line for data classification

while the super line's equation is $\frac{2}{\|w\|}$ (Ivanciuc 2007). In this study, 12 predictors and one predictable were available which added a deep complexity to the problem.

$$\vec{\mathbf{w}} \cdot \vec{\mathbf{x}} - b = 1 \tag{1}$$

$$\vec{\mathbf{w}} \cdot \vec{\mathbf{x}} - b = -1 \tag{2}$$

In pragmatic uses of SVM, the datasets commonly are in an *N*-dimensional space. Support vector machine linear machine of one output *y* (*x*), working in the high-dimensional feature space formed by the nonlinear mapping of *N*-dimensional input vector *x* into a *K*-dimensional feature space ($k > N$) with the nonlinear function $\varnothing(x)$. The number of hidden units or *K* is equal to the number of so-called support vector, that are learning data points, closest to the separating super line. The learning task transformed to the minimizing of the error function and simultaneously keeping the weights of the network at the possible minimum. The error function is defined through the so-called ε-insensitive loss function $L\varepsilon(d.y(x))$ (**Cortes and Vapnik 1995**).

$$L\varepsilon(d.y(x)) = \begin{cases} d - y(x) - \varepsilon & For\ (d - y(x)) \geq \varepsilon \\ 0 & For\ (d - y(x)) < \varepsilon \end{cases} \tag{3}$$

where ε supposed accuracy, *d* as destination, *x* as the input vector, and *y*(*x*) as the actual output signal of the SVM defined by:

$$y(x) = \sum_{j=1}^{K} W_j Q_j(x) + b = W^T \varnothing(x) + b \tag{4}$$

$w = [w_1. \ldots. w_K]^T$ is the weight vector, *b* represents bias, and $\varnothing(x) = [\varnothing_1. \ldots. \varnothing_K]^T$ the bias vector (Osowski and Garanty 2007). The solution of the so defined optimization problem solved by the introduction of the Lagrange multipliers $\alpha_i \alpha_i^*$ (where $i = 1.2. \ldots. K$) responsible for the functional constraints defined in Eq. (3). The minimization of the Lagrange function has been changed to the dual problem (Sapankevych and Sankar 2009):

$$\varnothing(\alpha.\alpha^*) = \left[\sum_{i=1}^{k} d_i(\alpha_i - \alpha_i^*) - \varepsilon\left(\sum_{i=1}^{k}(\alpha_i - \alpha_i^*) - \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k}(\alpha_i.\alpha_i^*)\left(\alpha_j.\alpha_j^*\right)K(x_i.x_j)\right]\right. \tag{5}$$

With constraints'

$$\sum_{i=1}^{k}(\alpha_i.\alpha_i^*) = 0$$

$$0 \leq \alpha_i \leq C \text{ and } 0 \leq \alpha_i^* \leq C$$

where *C* is a regularized constant that determines the tradeoff between the training risk and the model uniformity. According to the nature of quadratic programming, only those data corresponding to nonzero $(\alpha_i - \alpha_i^*)$ pairs can refer to support vectors (Nsv). In Eq. 5, $K(x_i.\ x_j) = \varnothing(x_i) \times \varnothing(x_j)$ is the inner product kernel which satisfies Mercer's condition (Schölkfopf et al. 1999) that is required for the generation of kernel functions given by:

$K(x_i.\ x_j) = \langle \varnothing(xi).\ \varnothing(X_j). \rangle$

Hence, the support vectors associates with the desired outputs y (*x*) and with the input training data *x* can define by

$$y(x) = \sum_{i=1}^{Nsv}(\alpha_i.\alpha_i^*) \cdot K(x.x_i) + b$$

Meteorological parameters such as average nebulosity, wind speed, sunshine, maximum, and minimum air temperature, relative humidity, and precipitation in addition to the chemical precursors like CO, SO_2, O_3, NO_2, and PM_{10} are building the variables and a simple linear classification is not able to categorize the datasets. In much complex problems, nonlinear vectors are required to classify (James et al. 2000). Kernel tricks transform the datasets into a *N*-dimensional space and then classify (Aronszajn 2009). With respect to the prior research about the kernel functions compatibilities' on a similar study (Mehdipour and Memarianfard 2017), the radial basis function (RBF) harnessed in the present paper. Meanwhile, the optimum amounts of sin2 and gamma have been revealed in the latest cited article; sin2 = 0.2 and gamma = 1. However, other kernel tricks such as linear kernel, polynomial (homogeneous and inhomogeneous), and hyperbolic tangent kernels have considerable potentials (Genton 2001; Theodoridis 2008). For prognosticating the $PM_{2.5}$ concentrations by the above-mentioned predictors, the 66% percent of the collected datasets used for training and the 15% allocated for the validation and the residual amount used for testing. In other words, from three consecutive years' data collection, two initial years' data allocated for machine training.

Bayesian network

Bayesian network was introduced by Bayes and Price (1763), a method belonging to a group of graphical probability modeling. Graphical structures employed to represent the information of a topic with uncertainty. Each node in a Bayesian graph shows a random variable and arcs or branches are depicting the probable relations between the variables where these conditional relations commonly are assessing by statistical tools (Varis and Kuikka 1999). Bayesian networks consist a combination of graph theory, probability theory, computer sciences, and statistics and have a wide utility in machine learning, data mining, sound identification, signal analyzing, bioinformatics, medical prognoses, and weather forecasting and specifically, there are numerous successful instances of Bayesian network application on environmental engineering (Vicedo-Cabrera et al. 2013; Uusitalo 2007; Wade 2000; Elizondo and Orun 2017; Nickless et al. 2017). The GeNIe 2.0 software has been employed in this study. Regarding collected datasets and their chemical and meteorological relations, as shown in Fig. 2, the arcs and their arrows have been set. Effects of all predictors on the $PM_{2.5}$ concentration as obligatory arcs, and relations between the predictors as random arcs opted for this study. Nonetheless, copious graphs and their compatibility have been analyzed and the best possible graph introduced. It is notable that, some arcs and arrows are merely statistically founded. As a tangible instance, wind speed (WS) has undeniable impacts on the humidity (H), particulate matter, nebulosity, etc.

Evaluation and comparison criteria

Root mean square error (RMSE) and correlation coefficient (CC) have been exploited in this research to assess the methods capabilities in producing and simulating the data which is akin to the test datasets. Equations 6 and 7 respectively representing the correlation coefficient and root mean square error. With respect to the recent equations, it can be achieved that lower RMSE (> 0) and higher CC (< 1) relates accuracy for the evolved models. Y_m and Y_p are the observed and predicted $PM_{2.5}$ and $\overline{y}m$ and $\overline{y}p$ are the average values for observed and simulated target variable. N shows the number of data for each parameter which is equal to the three consecutive years or 1096 days. CC and RMSE are the most reliable evaluation criteria (Chai and Draxler 2014; Roushangar and Homayounfar 2015) where have been used to compare the three above-mentioned methods. Also, Eq. 8 represents the normalized root mean square error (NRMSE) and Eq. 9 represents the Nash-Sutcliff coefficient (E). NRMSE is the non-dimensional form of RMSE and also the E coefficient can range from -∞ to 1 and $E = 1$ corresponds to a perfect match between the model and observations (Ömer Faruk 2010; Kuo et al. 2015; Lelieveld et al. 2015). X_{obs} and X_{model} are the observed and modeled values, respectively.

$$CC = \frac{\sum_{i=1}^{N}\left(Ym-\overline{y}m\right)\times\left(Yp-\overline{y}p\right)}{\sqrt{\sum_{i=1}^{N}\left(Ym-\overline{y}m\right)^{2}}\times\sqrt{\sum_{i=1}^{N}\left(Yp-\overline{y}p\right)^{2}}} \quad (6)$$

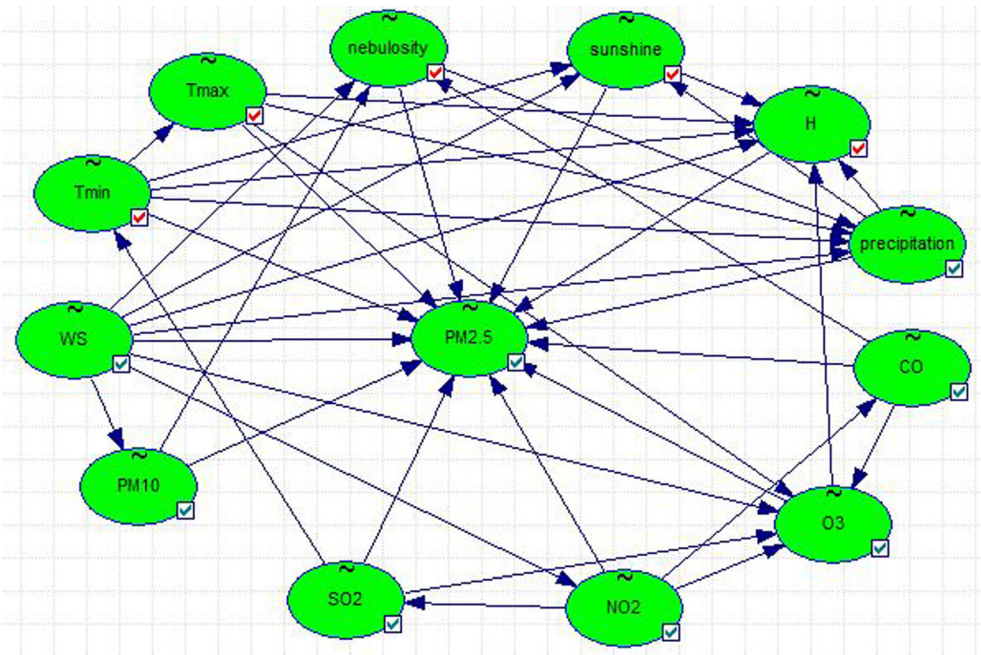

Fig. 2 Bayesian network of the predictors and predictable

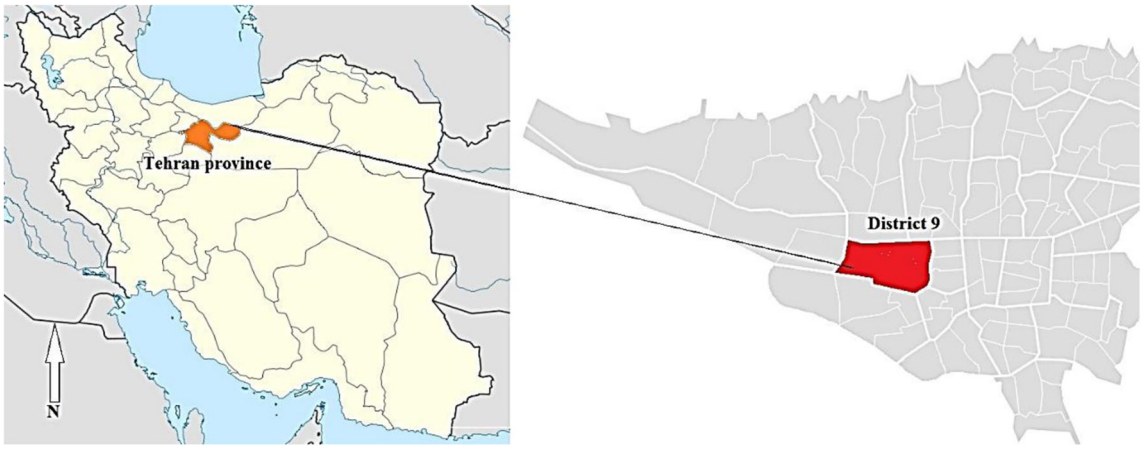

Fig. 3 The district 9 of Tehran county, Iran

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{N} \frac{(\mathrm{Ym}-\mathrm{Yp})^2}{N}} \tag{7}$$

$$NRMSE = \frac{\mathrm{RMSE}}{X_{\mathrm{obs,max}} - X_{\mathrm{obs,min}}} \tag{8}$$

$$E = 1 - \frac{\sum_{i=1}^{n} \left(X_{\mathrm{obs},i} - X_{\mathrm{mo\,del}}\right)^2}{\sum_{i=1}^{n} \left(X_{\mathrm{obs},i} - \overline{X_{\mathrm{obs}}}\right)^2} \tag{9}$$

Study area and datasets

Twenty ninth biggest metropolitan in the world is an unsecure nest for roughly 14 million residents during nights and 20 million commuter and resident on the daylight. An important industrial center in the heart of middle east plays the biggest role in Iran economy by possessing manifold factories. Factories placement near to residential area and their lack of facilities to reduce the air pollution is a detrimental factor for the people. Highly condensed traffic in Tehran's streets owing to the weak public transportation, crowded metros, expensive taxis, and other results grounds one the most dangerous air contamination for the Tehran people (Seyedabrishami and Mamdoohi 2012). The target study area has 1274 km^2 area and 22 municipal regions where located in 51° E longitude and 35° N latitude and 900 m up to 1830 m above the free seas altitude (Bagha et al. 2014). Each district has the air pollution measurement center; hence, 22 measuring centers hourly are gauging the contaminant concentration. $PM_{2.5}$, PM_{10}, CO, NO_2, SO_2, and O_3 are the measurable parameters. The meteorological parameters of Tehran are determined in district 9 where the Mehrabad airport is located. In this research, the parameters of the latest district have been employed. Figure 3 illustrates district 9 location.

Data collection and preparation

The datasets were collected from January 2013 to January 2016 for three consecutive years, 1096 days. The air pollution measuring station at district 9 gauges the air pollutants concentration every 3 h and in this paper, the maximum amount of every parameter was collected for each day. Furthermore, the meteorological variables were measured daily in Mehrabad

Table 1 Statistical descriptions of the input variables

	WS (km/h)	T_{min} (oC)	T_{max} (oC)	Neb (tenth)	Sunsh (lx)	RH (%)	Prec (mm)	CO (ppm)	O3 (ppm)	NO2 (ppm)	SO2 (ppm)	PM10 (μg/m^3)	PM2.5 (μg/m^3)
N	1096	1096	1096	1096	1096	1096	1096	1096	1096	1096	1096	1096	1096
Mean	3.2	13.7	23.6	2.6	8.2	33.7	0.4	39.2	33.4	58.1	26.1	61.5	93.3
Median	2.9	14.4	24.6	2.3	9.4	30.5	0.0	38.0	33.0	57.0	25.0	60.0	91.0
Mode	2.625	24	36	0	0	34	0	35	37	57	24	52	77
Std. deviation	1.47	9.09	10.65	2.11	3.60	17.12	1.78	9.53	11.93	11.15	5.30	18.94	26.31
Variance	2.16	82.70	113.44	4.46	12.92	292.32	3.16	89.57	141.36	121.25	27.52	355.45	684.71
Skewness	1.26	− 0.15	− 0.19	0.6	− 1.01	0.84	6.26	0.59	0.35	0.29	1.04	1.76	0.41
Kurtosis	2.09	− 1.15	− 1.20	− 0.61	− 0.08	0.01	47.46	0.16	− 0.27	− 0.26	1.36	12.36	− 0.05
Minimum	0.5	− 10.8	− 4.4	0.0	0.0	8.0	0.0	20.0	7.0	32.0	15.0	14.0	28.0
Maximum	10.1	32.6	42.6	8.0	13.5	88.4	21.0	77.0	74.0	97.0	50.0	252.0	190.0

Table 2 The correlation matrix of input datasets

Variables	WS	T_{min}	T_{max}	Nebulosity	Sunshine	RH	Prec	CO	O_3	NO_2	SO_2	PM_{10}	$PM_{2.5}$
WS	1.000	0.167	0.185	−0.036	0.168	−0.161	0.067	−0.289	0.172	−0.323	−0.215	−0.063	−0.239
T_{min}	0.167	1.000	0.972	−0.331	0.514	−0.637	−0.173	−0.036	0.715	−0.146	−0.333	0.124	−0.147
T_{max}	0.185	0.972	1.000	−0.320	0.520	−0.656	−0.173	−0.063	0.725	−0.168	−0.322	0.129	−0.159
Nebulosity	−0.036	−0.331	−0.320	1.000	−0.805	0.493	0.383	−0.221	−0.352	−0.184	−0.012	−0.174	−0.117
Sunshine	0.168	0.514	0.520	−0.805	1.000	−0.650	−0.376	0.048	0.527	0.046	−0.085	0.067	−0.084
RH	−0.161	−0.637	−0.656	0.493	−0.650	1.000	0.444	−0.122	−0.649	−0.072	0.045	−0.143	0.083
Prec	0.067	−0.173	−0.173	0.383	−0.376	0.444	1.000	−0.125	−0.183	−0.173	−0.131	−0.162	−0.129
CO	−0.289	−0.036	−0.063	−0.221	0.048	−0.122	−0.125	1.000	−0.070	0.700	0.229	0.296	0.445
O_3	0.172	0.715	0.725	−0.352	0.527	−0.649	−0.183	−0.070	1.000	−0.089	−0.093	0.098	−0.168
NO_2	−0.323	−0.146	−0.168	−0.184	0.046	−0.072	−0.173	0.700	−0.089	1.000	0.366	0.291	0.490
SO_2	−0.215	−0.333	−0.322	−0.012	−0.085	0.045	−0.131	0.229	−0.093	0.366	1.000	0.276	0.420
PM_{10}	−0.063	0.124	0.129	−0.174	0.067	−0.143	−0.162	0.296	0.098	0.291	0.276	1.000	0.848
$PM_{2.5}$	−0.239	−0.147	−0.159	−0.117	−0.084	0.083	−0.129	0.445	−0.168	0.490	0.420	0.848	1.000

airport. During the 3 years, Tehran experienced 35 clear and healthy air, 660 moderate air quality, 376 unhealthy for sensitive individuals, 24 unhealthy air pollution, and a day with very unhealthy air quality. Meanwhile, in 401 days, the $PM_{2.5}$ had the worst condition compared to the other pollutants. The parameters and values were gathered together from archives of Tehran air quality control company (http://airnow.tehran.ir) and meteorological organization of Iran (http://www.irimo.ir). Each of which is a reliable organization and equipped by updated apparatuses. Table 1 represents the statistical description of the collected datasets. WS, RH, Prec, T_{max}, T_{min}, Sunsh, and Neb are respectively the abbreviations of the wind velocity, relative humidity, precipitation, maximum temperature, minimum temperature, sunshine, and average nebulosity. Also, Table 2 represents the correlation matrix between the all collected parameters that illustrates which data has a positive or negative correlation with another. During data collection for this modeling study, there were undeniable obstacles and deficiencies. It is recommended to input other factors which may play a role in urban pollution in the future studies: daily fuel consumption, average number of commuters in the study area, traffic-related datasets, etc.

Equation 10 transfers the datasets in to a [0–1] limit to make the datasets comparable with each other. The monitored datasets have different units, e.g., the wind speed is measureable by kilometers per hour and the relative humidity is measuring by percentage; thus, data preparation is an indispensable step in this research. Data normalization makes it possible to have all parameters in a similar scale and more importantly to find a rational mathematical equation between

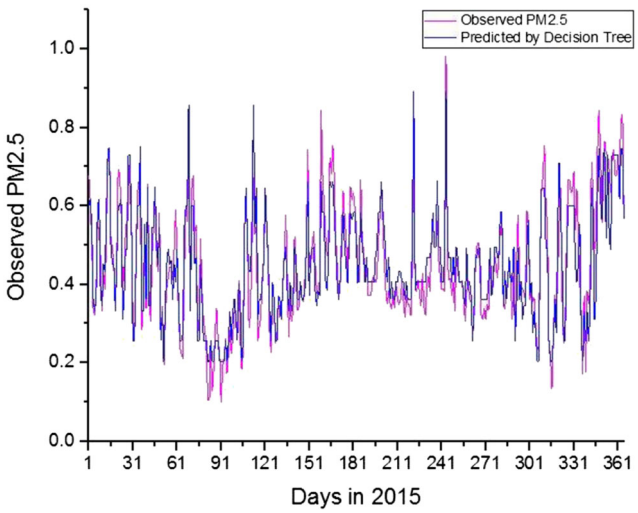

Fig. 4 Observed and predicted data by decision tree

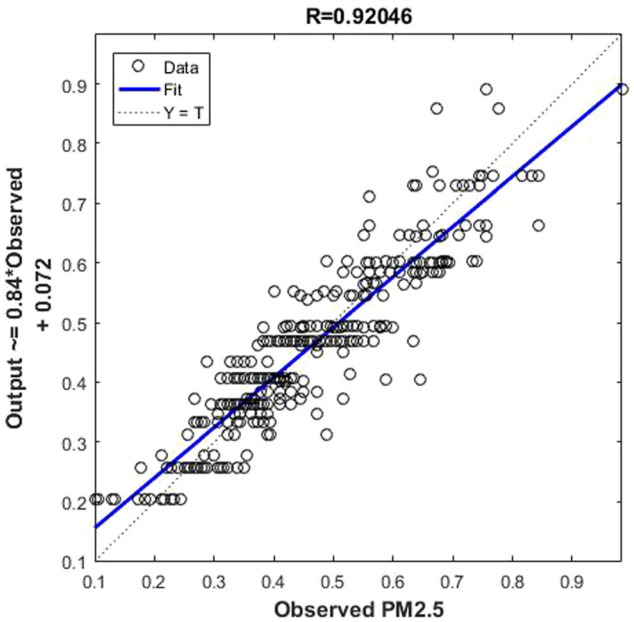

Fig. 5 The linear regression between the observed and predicted data by the decision tree

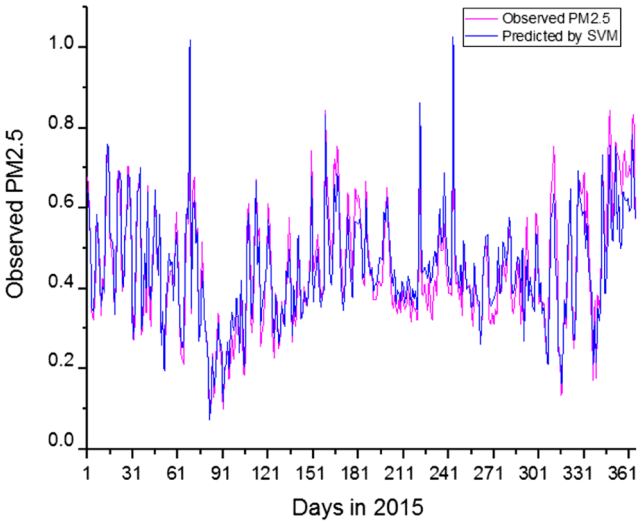

Fig. 6 Observed and predicted data by SVM

the predictors and predictable. $X_{\min}$ and $X_{\max}$ are the minimum and maximum of each variable and X_i represents the daily value of the parameters.

$$X = \frac{(X\mathrm{i} - X\mathrm{min})}{(\mathrm{Xmax} - \mathrm{Xmin})} \tag{10}$$

Results and discussions

In this paper, three modeling methods have been exploited to predict the $PM_{2.5}$ and each of which approaches abilities in simulating, showcased in this section to finally introduce the ablest tool. The most powerful method will be harnessed to exert the sensitivity analysis to measure the predictors' impacts on the variation of $PM_{2.5}$ concentration exerted.

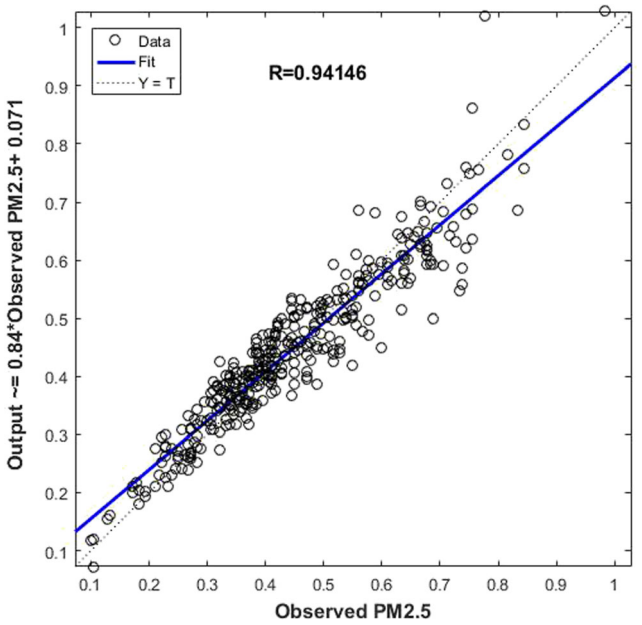

Fig. 7 Linear regression between the observed and predicted by SVM

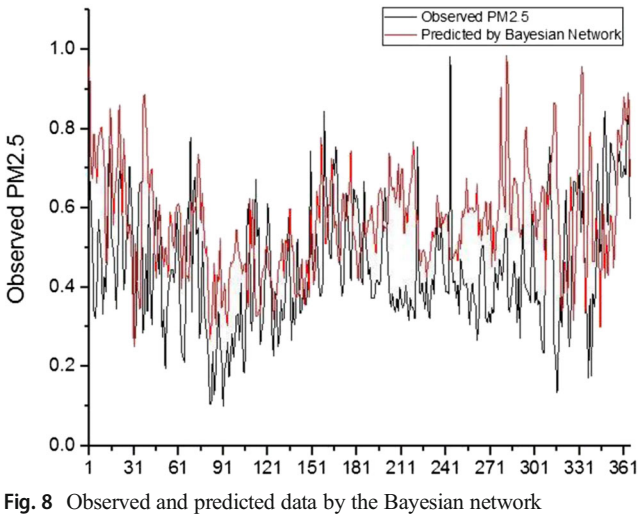

Fig. 8 Observed and predicted data by the Bayesian network

Results of the decision tree

Designed tree could provide acceptable results by generating a set of simulated data which compared to the observed $PM_{2.5}$ have RMSE equal to 0.0591. Furthermore, Figs. 4 and 5 respectively represent the linear regression for the evolved model and how the simulated datasets can follow the observed $PM_{2.5}$ in 2015. The correlation coefficient for the modeled data and observed data is 0.9204 which is in a quite acceptable range. The derived explicit equation from DT is provided in the Eqs. 11 and 12:

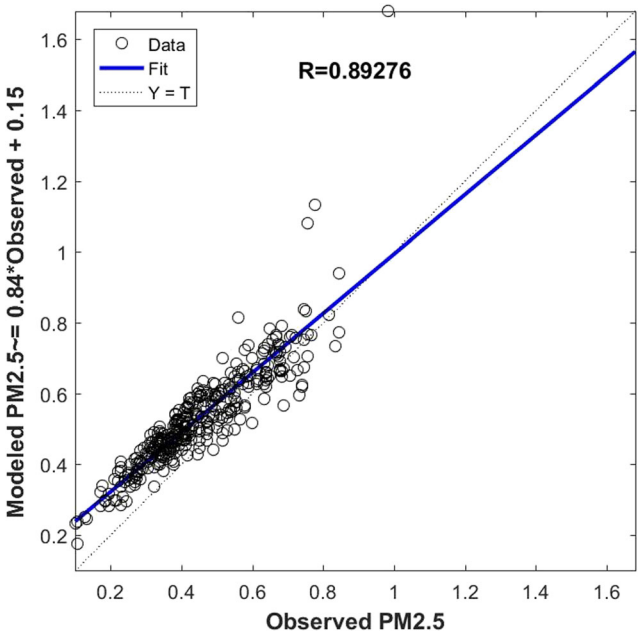

Fig. 9 Linear regression between the observed and predicted data by the Bayesian Network

Table 3 Evaluation criteria values of the developed models

	Decision tree	SVM	Bayesian network
RMSE	0.0591	0.0519	0.1077
CC	0.9204	0.9414	0.8927
NRMSE	0.0669	0.0587	0.1219
E	0.8472	0.8842	4943

If the $PM_{10} <= 0.291$, so

$$PM_{2.5} = -0.0294^{*}\,WS + 0.0359^{*}\,T_{min} - 0.0012^{*}\,T_{max} - 0.0218^{*}\,\text{nebulosity} + 0.0909^{*}\,RH + 0.0336^{*}\,CO - 0.0986^{*}\,O_3 + 0.1798^{*}\,NO_2 + 0.0613^{*}\,SO_2 + 1.7382^{*}\,PM_{10} - 0.1366 \quad (11)$$

But, if the $PM_{10} > 0.291$, so

$$PM_{2.5} = -0.0021^{*}\,WS + 0.0569^{*}\,T_{min} - 0.0785^{*}\,T_{max} - 0.0648^{*}\,\text{nebulosity} - 0.0474^{*}\,\text{sunshine} + 0.1601^{*}RH - 0.2033^{*}\,\text{precipitation} - 0.1118^{*}\,O_3 + 0.17^{*}\,NO_2 + 0.0601^{*}\,SO_2 + 1.2992^{*}\,PM_{10} + 0.1146 \quad (12)$$

Results of support vector machine

Figure 6 illustrates the predicted and simulated values of $PM_{2.5}$ in one graph to show that outputs of the built model roughly follow the observed datasets. Also, Fig. 7 represents the linear regression between the observed and predicted $PM_{2.5}$ simulated by the support vector machine which shows a quite acceptable result as the correlation coefficient is equal to 0.9414. Overfitting is a menace for soft computing methods which harm the models' accuracy and this happens when the results for test data are better than the result for the train datasets. However, in this study, over-training or overfitting is controlled, as the CC and RMSE for the training datasets respectively are 0.9426 and 0.0501. Root mean square error for the testing datasets and produced datasets is 0.0519. Thus, the support vector machine is not over-trained.

Results of Bayesian network

In this study, all predictors' effects on $PM_{2.5}$ concentration have been considered. Simultaneously, the predictors have relations with each other and in the present Bayesian Network structure, their relations went under study to have a more accurate structure (see Fig. 2). For estimation of $PM_{2.5}$, the Bayesian network gave a function considering all parameters which is shown in the Eq. 13, where the WS, T_{min}, T_{max}, N, S, H, P, CO, O_3, NO_2, SO_2, and PM_{10} represent the wind speed, daily minimum temperature, maximum temperature of the day, nebulosity, sunshine, relative humidity, participation, carbon dioxide, ground level ozone, nitrogen dioxide, sulfur dioxide, and particulate matters with 10-μm diameters, respectively.

$$PM_{2.5} = -0.041 \times WS + 0.055 \times T_{min} - 0.027 \times T_{max} - 0.032 \times N - 0.011 \times S + 0.093 \times RH - 0.028 \times P + 0.021 \times CO - 0.133 * O_3 + 0.197 \times NO_2 + 0.101 \times SO_2 + 1.616 \times PM_{10} \quad (13)$$

By exploiting MS Excel software and Eq. 13, the modeled $PM_{2.5}$ values produced. Figure 8 shows how simulated data follow the test data. The RMSE value between the modeled $PM_{2.5}$ by the BN and observed is equal to 0.1077, and as shown in Fig. 9, the correlation coefficient is 0.8927.

Comparing the methods

The RMSE, NRMSE, CC, and E of each modeled data for the testing datasets have been assessed by four most prominent evaluation criteria. All three methods gave acceptable results in various fields of study. Evolved models are easily comparable regarding Table 3. Further, single factor analysis of variance (ANOVA) tested to compare their robustness of methods with each other (Sihag et al. 2018a, b). Table 4 shows that DT and SVM have an *F* value less than F critical and the *P* values for these two methods are greater than 0.05, while the *F* value for BN is more than critical amount and also the *P* value of BN is less than 0.05; therefore, the DT and SVM are unbiased methods and their predicted values are insignificantly different from observed data. On the other hand, the BN is biased and results of estimated and actual amount are significantly different.

According to Tables 3 and 4, SVM yielded a meaningful power in comparison to the other methods in this study and other studies of the writers; hence, application of this modeling system and combining it with other

Table 4 The single factor ANOVA for methods

Approaches	*F*	*P* value	*F* crit	Difference between actual and predicted values
DT	1.86714E-05	0.996553498	3.854263749	Insignificant
SVM	0.003702706	0.951495479	3.854263749	Insignificant
BN	58.05200233	7.96038E-14	3.854263749	Significant

Table 5 $PM_{2.5}$ sensitivity analyses' results for different input combinations by SVM

Model	Components of each model	RMSE	R	NRMSE	E
SVM01	T_{max}	0.1026	0.6864	0.1161	0.3369
SVM02	T_{max}, WS	0.1016	0.6880	0.1150	0.3436
SVM03	T_{max}, WS, S	0.1040	0.6883	0.1177	0.3965
SVM04	T_{max}, WS, S, T_{min}	0.0999	0.6899	0.1131	0.4040
SVM05	T_{max}, WS, S, T_{min}, Neb	0.0892	0.6950	0.1010	0.4614
SVM06	T_{max}, WS, S, T_{min}, Neb, RH	0.0875	0.7130	0.0990	0.5053
SVM07	T_{max}, WS, S, T_{min}, N, RH, Prec	0.0860	0.7610	0.0974	0.5627
SVM08	T_{max}, WS, S, T_{min}, N, RH, Prec, CO	0.0849	0.7710	0.0961	0.6033
SVM09	T_{max}, WS, S, T_{min}, N, RH, Prec, CO, O_3	0.0840	0.8352	0.0951	0.6522
SVM10	T_{max}, WS, S, T_{min}, N, RH, Prec, CO, O_3, NO_2	0.0790	0.8933	0.0894	0.7452
SVM11	T_{max}, WS, S, T_{min}, N, RH, Prec, CO, O_3, NO_2, SO_2	0.0770	0.9003	0.0872	0.8064
SVM12	T_{max}, WS, S, T_{min}, N, RH, Prec, CO, O_3, NO_2, SO_2, PM_{10}	0.0519	0.9414	0.0587	0.8842

possible methods is strongly suggested. Specifically, a hybrid of least square and support vector machine or LSSVM anticipated to produce potent models. Respectively DT and BN are the in next places.

Sensitivity analysis of $PM_{2.5}$ via SVM

$PM_{2.5}$ sensitivity analysis against all of predictors is depicted in Table 5. SVM as the ablest method of this research is selected to run sensitivity analysis. According to the latest studies about the capability of different Kernel functions, the radial basis function or RBF has been chosen as the Kernel trick of the SVM (Mehdipour and Memarianfard 2018; Sihag et al. 2018a, b). In this analysis, predictor parameters added one by one and the model ran for each input variable. Finally, effects of each parameter on $PM_{2.5}$ tolerances can be detected by comparison the RMSE, NRMSE, CC, and *E* values. Model SVM12 has the optimum results.

Conclusion

Air pollution measuring instruments are expensive, massive, and hardly maintainable. Thus, a reliable soft method can be a proper substitute. For this aim, Bayesian network (BN), decision tree (DT), and support vector machine (SVM) applied to model $PM_{2.5}$ concentration. Regarding the evaluation criteria, SVM introduced as the ablest method and DT and BN are in the next places.

With respect to the provided mathematical equations by BN and DT, and sensitivity analysis of $PM_{2.5}$ via SVM, the predictors effects are comprehensible; highly effective parameters have a higher coefficient in the suggested equations by BN or DT and vice versa. Also, adding parameters with a higher influence can reduce the RMSE or NRMSE and escalate the CC or *E* values more than others, in the sensitivity analysis table. PM_{10} has the greatest impact on the prediction of the $PM_{2.5}$ and chemical precursors have more influences on the $PM_{2.5}$ variances in comparison to meteorological parameters. However, as the particulate matters are prone to adhesion and subsiding along with the humidity, it influences the $PM_{2.5}$ significantly. Also, wind speed was anticipated to have a higher impact, as wind can carry the particulate matter, but in this study, variances of the wind velocity does not undeniably effect the $PM_{2.5}$ value. Authors suggest to study on the wind speed and possible reasons of its low effects on the particulate matters; however, it is postulated that besieging the city by skyscrapers and low values of wind speed are the main reasons.

Acknowledgements This article is in memories of professor S. A. Sadrnejad whom we missed with great regrets. Also, the authors are grateful to dear Farzin Homayounfar, Dr. E. Kouhestani, and other collaborators who suggested their invaluable comments.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interests.

References

Aguilera PA, Fernández A, Fernández R, Rumí R, Salmerón A (2011) Bayesian networks in environmental modelling. Environ Model Softw 26:1376–1388. https://doi.org/10.1016/j.envsoft.2011.06.004

Aronszajn N (2009) Theory of reproducing kernels. Am Math Soc 68: 337–404. https://doi.org/10.2307/1990404

Atkinson RW, Kang S, Anderson HR, Mills IC, Walton HA (2014) Epidemiological time series studies of PM2.5 and daily mortality and hospital admissions: a systematic review and meta-analysis. Thorax 69:660–665. https://doi.org/10.1136/thoraxjnl-2013-204492

Bagha N, Arian M, Ghorashi M, Pourkermani M, el Hamdouni R, Solgi A (2014) Geomorphology evaluation of relative tectonic activity in

the Tehran basin, central Alborz, northern Iran. Geomorphology 29: 135–145. https://doi.org/10.1016/j.geomorph.2013.12.041

Bayes M, Price M (1763) An essay towards solving a problem in the doctrine of chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. Philos Trans R Soc Lond 53:370–418. https://doi.org/10.1098/rstl.1763.0053

Borja-Aburto VH, Castillejos M, Gold DR, Bierzwinski S, Loomis D (1998) Mortality and ambient fine particles in southwest Mexico City, 1993-1995. Environ Health Perspect 106:849–855. https://doi.org/10.2307/3434129

Cao J, Chow JC, Lee FSC, Watson JG (2013) Evolution of PM2.5 measurements and standards in the U.S. and future perspectives for China. Aerosol Air Qual Res 13:1197–1211. https://doi.org/10.4209/aaqr.2012.11.0302

Chai T, Draxler RR (2014) Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. Geosci Model Dev 7:1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Cortes C, Vapnik V (1995) Support vector networks. Mach Learn 20: 273–297. https://doi.org/10.1007/BF00994018

Davidson CI, Phalen RF, Solomon PA (2005) Airborne particulate matter and human health: a review. Aerosol Sci Technol 39:737–749. https://doi.org/10.1080/02786820500191348

Dunea D, Iordache S, Liu H-Y, Bøhler T, Pohoata A, Radulescu C (2016) Quantifying the impact of PM2.5 and associated heavy metals on respiratory health of children near metallurgical facilities. Environ Sci Pollut Res Int 23:15395–15406. https://doi.org/10.1007/s11356-016-6734-x

Elizondo D, Orun A (2017) An Intelligent traffic network optimisation by use of Bayesian inference methods to combat air pollution. In: TPM-Transport Practitioner's Meeting Conference, 28–29 June 2017, Nottingham

Fann N, Lamson AD, Anenberg SC, Wesson K, Risley D, Hubbell BJ (2012) Estimating the national public health burden associated with exposure to ambient PM2.5 and ozone. Risk Anal 32:81–95. https://doi.org/10.1111/j.1539-6924.2011.01630.x

Fattore E, Paiano V, Borgini A, Tittarelli A, Bertoldi M, Crosignani P, Fanelli R (2011) Human health risk in relation to air quality in two municipalities in an industrialized area of northern Italy. Environ Res 111:1321–1327. https://doi.org/10.1016/j.envres.2011.06.012

Feng X, Li Q, Zhu Y, Hou J, Jin L, Wang J (2015) Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation. Atmos Environ 107:118–128. https://doi.org/10.1016/j.atmosenv.2015.02.030

Genton MG (2001) Classes of kernels for machine learning: a statistics perspective. J Mach Learn Res 2:299–312. https://doi.org/10.1162/15324430260185646

Ivanciuc O (2007) Applications of support vector machines in chemistry. In: Reviews in Computational Chemistry. pp 291–400

James G, Witten D, Hastie T, Tibshirani R (2000) An introduction to statistical learning. Springer New York Heidelberg Dordrecht London

Kamiński B, Jakubczyk M, Szufel P (2017) A framework for sensitivity analysis of decision trees. Cent Eur J Oper Res 26:1–25. https://doi.org/10.1007/s10100-017-0479-6

Kim S, Shiri J, Singh VP, Kisi O, Landeras G (2015) Predicting daily pan evaporation by soft computing models with limited climatic data. Hydrol Sci J 60:1120–1136. https://doi.org/10.1080/02626667.2014.945937

Kisi O, Parmar KS, Soni K, Demir V (2017) Modeling of air pollutants using least square support vector regression, multivariate adaptive regression spline, and M5 model tree models. Air Qual Atmos Health 10:873–883. https://doi.org/10.1007/s11869-017-0477-9

Kujaroentavon K, Kiattisin S, Leelasantitham A, Thammaboosadee S (2015) Air quality classification in Thailand based on decision tree. In: BMEiCON 2014-7th Biomedical Engineering International Conference

Kuo Y-M, Chiu C-H, Yu H-L (2015) Influences of ambient air pollutants and meteorological conditions on ozone variations in Kaohsiung, Taiwan. Stoch Env Res Risk A 29:1037–1050. https://doi.org/10.1007/s00477-014-0968-2

Leili M, Naddafi K, Nabizadeh R, Yunesian M, Mesdaghinia A (2008) The study of TSP and PM10 concentration and their heavy metal content in central area of Tehran, Iran. Air Qual Atmos Health 1: 159–166. https://doi.org/10.1007/s11869-008-0021-z

Lelieveld J, Evans JS, Fnais M, Giannadaki D, Pozzer A (2015) The contribution of outdoor air pollution sources to premature mortality on a global scale. Nature 525:367–371

Li Q, Guo Y, Song J-Y, Song Y, Ma J, Wang HJ (2018) Impact of long-term exposure to local PM10 on children's blood pressure: a Chinese national cross-sectional study. Air Qual Atmos Health 11: 705–713. https://doi.org/10.1007/s11869-018-0577-1

Liu JC, Peng RD (2018) Health effect of mixtures of ozone, nitrogen dioxide, and fine particulates in 85 US counties. Air Qual Atmos Health 11:311–324. https://doi.org/10.1007/s11869-017-0544-2

Liu KF-R, Lu C-F, Chen C-W, Shen Y-S (2012) Applying Bayesian belief networks to health risk assessment. Stoch Env Res Risk A 26:451–465. https://doi.org/10.1007/s00477-011-0470-z

Marchant R, Ramos F (2012) Bayesian optimisation for intelligent environmental monitoring. IEEE Int Conf Intell Robot Syst 2242–2249. doi: https://doi.org/10.1109/IROS.2012.6385653

Martí P, Shiri J, Duran-ros M, et al (2013) Artificial neural networks vs. Gene Expression Programming for estimating outlet dissolved oxygen in micro-irrigation sand filters fed with effluents 99:176–185. doi: https://doi.org/10.1016/j.compag.2013.08.016

Marzouni MB, Alizadeh T, Banafsheh MR, Khorshiddoust AM, Ghozikali MG, Akbaripoor S, Sharifi R, Goudarzi G (2016) A comparison of health impacts assessment for PM10 during two successive years in the ambient air of Kermanshah, Iran. Atmos Pollut Res 7:1–7. https://doi.org/10.1016/j.apr.2016.04.004

McCann RK, Marcot BG, Ellis R (2006) Bayesian belief networks: applications in ecology and natural resource management. Can J For Res 36:3053–3062. https://doi.org/10.1139/x06-238

McMillan NJ, Holland DM, Morara M, Jingyu F (2007) Space time zero inflated count models of harbor seals. Environmetrics 18:697–712. https://doi.org/10.1002/env

Mehdipour V (2017) Temporal modeling of tropospheric ozone and analysis of its relationship with photochemical precursors considering meteorological parameters. K. N. Toosi University of Technology, Tehran

Mehdipour V, Memarianfard M (2017) Application of support vector machine and gene expression programming on tropospheric ozone prognosticating for Tehran metropolitan. Civ Eng J 3:557. https://doi.org/10.28991/cej-030984

Mehdipour V, Memarianfard M (2018) Ground-level O3 sensitivity analysis using support vector machine with radial basis function. Int J Environ Sci Technol. https://doi.org/10.1007/s13762-018-1770-3

Mehdipour V, Memarianfard M, Homayounfar F (2017) Application of gene expression programming to water dissolved oxygen concentration prediction. International Journal of Human Capital in Urban Management 2:39–48. https://doi.org/10.22034/ijhcum.2017.02.01.004

Moret BME (1982) Decision trees and diagrams. ACM Comput Surv 14: 593–623. https://doi.org/10.1145/356893.356898

Nickless A, Rayner PJ, Engelbrecht F, Brunke EG, Erni B, Scholes RJ (2017) Estimates of CO 2 fluxes over the City of Cape Town, South Africa, through Bayesian inverse modelling. Atmos Chem Phys :1–72. https://doi.org/10.5194/acp-2017-604

Ömer Faruk D (2010) A hybrid neural network and ARIMA model for water quality time series prediction. Eng Appl Artif Intell 23:586–594. https://doi.org/10.1016/j.engappai.2009.09.015

Osowski S, Garanty K (2007) Forecasting of the daily meteorological pollution using wavelets and support vector machine. Eng Appl Artif Intell 20:745–755. https://doi.org/10.1016/j.engappai.2006.10.008

Pascal M, Corso M, Chanel O, Declercq C, Badaloni C, Cesaroni G, Henschel S, Meister K, Haluza D, Martin-Olmedo P, Medina S, Aphekom group (2013) Assessing the public health impacts of urban air pollution in 25 European cities: results of the Aphekom project. Sci Total Environ 449:390–400. https://doi.org/10.1016/j.scitotenv.2013.01.077

Pope CA, Ezzati M, Cannon JB, Allen RT, Jerrett M, Burnett RT (2018) Mortality risk and PM2.5 air pollution in the USA: an analysis of a national prospective cohort. Air Qual Atmos Health 11:245–252. https://doi.org/10.1007/s11869-017-0535-3

Quinlan JR (2006) Simplifying decision trees. Int J:221–234

Rivest RL (1987) Learning decision lists. Mach Learn 2:229–246. https://doi.org/10.1023/A:1022607331053

Roushangar K, Homayounfar F (2015) Prediction of flow friction coefficient using GEP and ANN methods. International Journal of Artificial Intelligence and Mechatronics 4:65–68

Sapankevych N, Sankar R (2009) Time series prediction using support vector machines: a survey. IEEE Comput Intell Mag 4:24–38. https://doi.org/10.1109/MCI.2009.932254

Schölkfopf B, Smola AJ, Burges C (1999) Advances in kernel methods: support vector learning. MIT Press, London

Schweitzer L, Zhou J (2010) Neighborhood air quality, respiratory health, and vulnerable populations in compact and sprawled regions. J Am Plan Assoc 76:363–371. https://doi.org/10.1080/01944363.2010.486623

Seyedabrishami S, Mamdoohi A (2012) Impact of carpooling on fuel saving in urban transportation: case study of Tehran. Procedia Soc Behav Sci 54:323–331. https://doi.org/10.1016/j.sbspro.2012.09.751

Sfetsos A, Vlachogiannis D (2010) A new approach to discovering the causal relationship between meteorological patterns and PM10 exceedances. Atmos Res 98:500–511. https://doi.org/10.1016/j.atmosres.2010.08.021

Sharifi SS, Rezaverdinejad V, Nourani V (2016) Estimation of daily global solar radiation using wavelet regression, ANN, GEP and empirical models: a comparative study of selected temperature-based approaches. J Atmos Sol Terr Phys 149:131–145. https://doi.org/10.1016/j.jastp.2016.10.008

Sihag P, Jain P, Kumar M (2018a) Modelling of impact of water quality on recharging rate of storm water filter system using various kernel function based regression. Modeling Earth Systems and Environment 4:61–68. https://doi.org/10.1007/s40808-017-0410-0

Sihag P, Singh B, Vand AS, Mehdipour V (2018b) Modeling the infiltration process with soft computing techniques. ISH Journal of Hydraulic Engineering 5010:1–15. https://doi.org/10.1080/09715010.2018.1464408

Theodoridis S (2008) Pattern recognition, 4th editio. Academic, Burlington

Utgoff PE (1989) Incremental induction of decision trees. Mach Learn 4:161–186. https://doi.org/10.1023/A:1022699900025

Uusitalo L (2007) Advantages and challenges of Bayesian networks in environmental modelling. Ecol Model 203:312–318. https://doi.org/10.1016/j.ecolmodel.2006.11.033

Vafa-arani H, Jahani S, Dashti H et al (2014) A system dynamics modeling for urban air pollution: a case study of Tehran, Iran. Transp Res Part D: Transp Environ 31:21–36. https://doi.org/10.1016/j.trd.2014.05.016

Varis O, Kuikka S (1999) Learning Bayesian decision analysis by doing: lessons from environmental and natural resources management. Ecol Model 119:177–195. https://doi.org/10.1016/S0304-3800(99)00061-7

Vicedo-Cabrera AM, Biggeri A, Grisotto L, Barbone F, Catelan D (2013) A Bayesian kriging model for estimating residential exposure to air pollution of children living in a high-risk area in Italy. Geospat Health 8:87–95. https://doi.org/10.4081/gh.2013.57

Wade PR (2000) Bayesian methods in conservation biology. Conserv Biol 14:1308–1316. https://doi.org/10.1046/j.1523-1739.2000.99415.x

Wang P, Liu Y, Qin Z, Zhang G (2015) Science of the total environment a novel hybrid forecasting model for PM 10 and SO 2 daily concentrations. Sci Total Environ 505:1202–1212. https://doi.org/10.1016/j.scitotenv.2014.10.078

World Health Organization (2003) Health aspects of air pollution with particulate matter, ozone and nitrogen dioxide: report on a WHO working group, Bonn, Germany 13–15 January 2003

Xing YF, Xu YH, Shi MH, Lian YX (2016) The impact of PM2.5 on the human respiratory system. J Thorac Dis 8:E69–E74. https://doi.org/10.3978/j.issn.2072-1439.2016.01.19