Science of the Total Environment 701 (2020) 134474

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Spatial hazard assessment of the PM10 using machine learning models in Barcelona, Spain

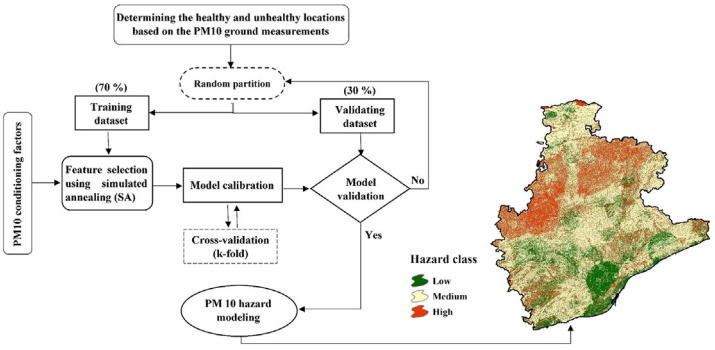
Bahram Choubin [a], Mahsa Abdolshahnejad [b], Ehsan Moradi [b], Xavier Querol [c], Amir Mosavi [d,e], Shahaboddin Shamshirband [f,*], Pedram Ghamisi [g]

[a] Soil Conservation and Watershed Management Research Department, West Azarbaijan Agricultural and Natural Resources Research and Education Center, AREEO, Urmia, Iran
[b] Department of Reclamation of Arid and Mountainous Regions, Faculty of Natural Resources, University of Tehran, Karaj, Iran
[c] Institute of Environmental Assessment and Water Research (IDAEA), Spanish Council for Scientific Research (CSIC), Barcelona, Spain
[d] School of the Built Environment, Oxford Brookes University, Oxford, UK
[e] Institute of Automation, Kando Kalman Faculty of Electrical Engineering, Obuda University, 1034 Budapest, Hungary
[f] Institute of Research and Development, Duy Tan University, Da Nang 550000, Viet Nam
[g] Exploration Devision, Helmholtz Institute Freiberg for Resource Technology, Helmholtz-Zentrum Dresden-Rossendorf, Dresden, Germany

HIGHLIGHTS

- Hazard prediction of particulate matter (PM) by the Machine learning (ML) models.
- Selection of key features using the simulated annealing (SA) method.
- Good performance of the ML models in PM modeling (Accuracy > 87%; Precision > 86%).

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 10 August 2019
Received in revised form 13 September 2019
Accepted 14 September 2019
Available online 4 October 2019

Editor: Pavlos Kassomenos

Keywords:
Hazard assessment
Particulate matter
Air quality
Simulated annealing
Random forest
Bagged classification and regression trees
Mixture discriminate analysis

ABSTRACT

Air pollution, and especially atmospheric particulate matter (PM), has a profound impact on human mortality and morbidity, environment, and ecological system. Accordingly, it is very relevant predicting air quality. Although the application of the machine learning (ML) models for predicting air quality parameters, such as PM concentrations, has been evaluated in previous studies, those on the spatial hazard modeling of them are very limited. Due to the high potential of the ML models, the spatial modeling of PM can help managers to identify the pollution hotspots. Accordingly, this study aims at developing new ML models, such as Random Forest (RF), Bagged Classification and Regression Trees (Bagged CART), and Mixture Discriminate Analysis (MDA) for the hazard prediction of PM10 (particles with a diameter less than 10 μm) in the Barcelona Province, Spain. According to the annual PM10 concentration in 75 stations, the healthy and unhealthy locations are determined, and a ratio 70/30 (53/22 stations) is applied for calibrating and validating the ML models to predict the most hazardous areas for PM10. In order to identify the influential variables of PM modeling, the simulated annealing (SA) feature selection method is used. Seven features, among the thirteen features, are selected as critical features. According to the results, all the three-machine learning (ML) models achieve an excellent performance (Accuracy > 87% and precision > 86%). However, the Bagged CART and RF models have the same performance and higher than the

* Corresponding author.
E-mail address: shamshirbandshahaboddin@duytan.edu.vn (S. Shamshirband).

https://doi.org/10.1016/j.scitotenv.2019.134474

MDA model. Spatial hazard maps predicted by the three models indicate that the high hazardous areas are located in the middle of the Barcelona Province more than in the Barcelona's Metropolitan Area.

1. Introduction

During the last decades, the rapid industrialization, increasing the fossil fuel consumption, population growth, urbanization, and other human-induced changes have dramatically increased air pollution worldwide (Daniel et al., 2009; Dall'Osto et al., 2012; Liu et al., 2014). The air pollution and particulate matter (PM) have a broad negative effect on the human health, mortality, morbidity, as well as the environment and ecological systems (Donaldson and MaCnee, 1999; Hoek et al., 2002a,b; Brunelli et al., 2007; Chen and Kan, 2008; Franklin et al., 2008; Peng et al., 2009; Fortelli et al., 2016). The inhalable atmospheric particulate matter (PM) with a diameter finer than 10 μm are referred to as PM10, which are responsible for large health worldwide impacts (Ganguly et al. 2019; Feng et al. 2019). The epidemiological researches demonstrated the long-term and short-term health impacts of PM10 and also finer PM on public health (Abbey et al., 1999; Pope et al., 2002; Brunekreef and Holgate, 2002; Dominici et al., 2006; Beelen et al., 2014; Brook et al., 2010; Rückerl et al., 2011; Stafoggia et al., 2013). For instance, according to the World Health Organization (WHO), PM pollution, only in Spain, is annually responsible for 25,000 premature deaths (Fernandez-Navarro et al., 2017; García Nieto et al., 2018). Health concern about PM10 has dramatically increased where the research has linked PM10 to significant health issues (Marchetti et al., 2019; Liu et al. 2019). The WHO, as well as many cancer research institutes, identify the PM as carcinogenic and highly harmful to humans (Dehghan et al., 2018). All these concerns accounted for a large increase of studies aiming at simulating and predicting PM pollution (Luo et al., 2018).

There is a wide range of methods to simulate and predict PM pollution. For instance, deterministic methods are widely used, e.g., community multi-scale air quality (Chen et al., 2013; Djalalova et al., 2015). These methods generally use coupled atmospheric transport, emission and chemical modules (i.e. Zhou et al., 2019). The validity of these methods depends on the availability, quality and scale of data, and also in order to predict through such processes, the methods need very large efforts (Zhang et al., 2012; Niu et al., 2016). However, to perform sensitivity analysis to assess policy measures, these are the only available tools.

Another approach for predicting air quality (including PM levels) is based on statistical modelling, e.g., mixed-effects models and linear regressions (Chen et al., 2018a,b,c; Gupta and Christopher, 2009; Lee et al., 2011; Liu et al., 2009). In this regard, some studies have indicated that the statistical methods such as linear regression models can increase the accuracy of the results of deterministic methods (Konovalov et al., 2009; Song et al., 2015). Although the result of using the traditional and parametric regression model sometimes has a high validity for prediction, these models cannot wholly obtain the relation between predictors and PM levels, also, they faced many problems with various data sets and predictors (Kloog et al., 2014; Hu et al., 2017; Chen et al., 2018a,b,c). Therefore, both of the deterministic and statistical methods cannot correctly simulate the complex issues (Choubin et al., 2019a,b).

In the Mediterranean basin, not only the specific climatic, geographic and anthropogenic emissions, but also the long range transport of desert dust cause relatively high levels of PM in specific areas (Escudero et al., 2007; Amodio et al., 2008; Querol et al., 2009, among others). Key factors, such as volume of emissions, time of the day, temperature, insolation, humidity, precipitation, wind speed, and direction govern changes in PM concentration (Rodriguez et al., 2001; Amodio et al., 2008; De Gennaro et al., 2013, among others). Therefore, it is complex to reach a high PM forecast accuracy due to the complexity of the system (Feng et al., 2015; Liu and Li, 2015; Luo et al., 2018). Recently, using advanced statistical methods, such as artificial intelligence (AI) and machine learning (ML) models for air pollution modelling and prediction have become popular (Lei et al., 2009; Kumar and Ridder, 2010; Bai et al. 2018; Delavar et al. 2019, among others). The foundation of the ML models is related to finding optimized algorithms based on computational statistics (Chen et al., 2018a, b,c). Chen et al. (2018b) utilized the multi-layer perceptron (MLP) network to simulate the PM10 concentrations. In another study, Chen et al. (2018a) compared the ability of ML methods with traditional regression models to evaluate the PM contaminations. The results showed that the ML models achieved more accuracy in prediction. De Gennaro et al. (2013) evaluated the application of artificial neural networks for the prediction of PM10 and evidenced that this method had a good ability for prediction. However, there are some studies on the use of ML models to predict air quality, including PM levels (Taspinar, 2015; Lary et al., 2015; Gardner and Dorling's, 1998; Corani, 2005; Gupta and Christopher, 2009; Voukantsis et al, 2011; Elangasinghe et al., 2014; Oprea et al., 2016; Zhu et al., 2018; Suleiman et al., 2019; Nabavi et al., 2019). Although ML methods demand more observation data for learning, they are faster and more efficient than the traditional ones. Besides, they include fewer limitation by some assumptions (Wang et al., 2017; Hu et al., 2017). In fact, due to the learning potential of the ML models, the relationship between air pollutants (such as PM) and predictors in different atmospheric conditions can be considered by applying them (Cobourn, 2010; Hrust et al., 2009).

Due to the high potential of the ML models, the spatial modeling of PM levels can help managers to identify the most hazardous areas. Hence this study aims to develop new ML models such as Random Forest (RF), Bagged Classification and Regression Trees (Bagged CART), and Mixture Discriminate Analysis (MDA) for PM10 hazard prediction. Although the application of the RF model for predicting the air quality parameters, including PM concentrations have been evaluated in prior studies (such as Hu et al., 2017; Chen et al., 2018a,b,c; Stafoggia et al., 2019), those on the spatial modeling of PM10 is very limited. Thus, this is a young research area, and, in this context, the contribution of the Bagged CART and MDA models is novel. The main objectives of the study are: (i) identification of the critical variables on PM modeling through the simulated annealing (SA) feature selection method, (ii) prediction of spatial hazard maps for PM10 pollution, and (iii) comparison of the novel ML models in prediction of PM hazard.

2. Material and methods

2.1. Study area

The Barcelona Province, in NE Spain was selected as the study region in our study, which extends between latitudes of 41°15′

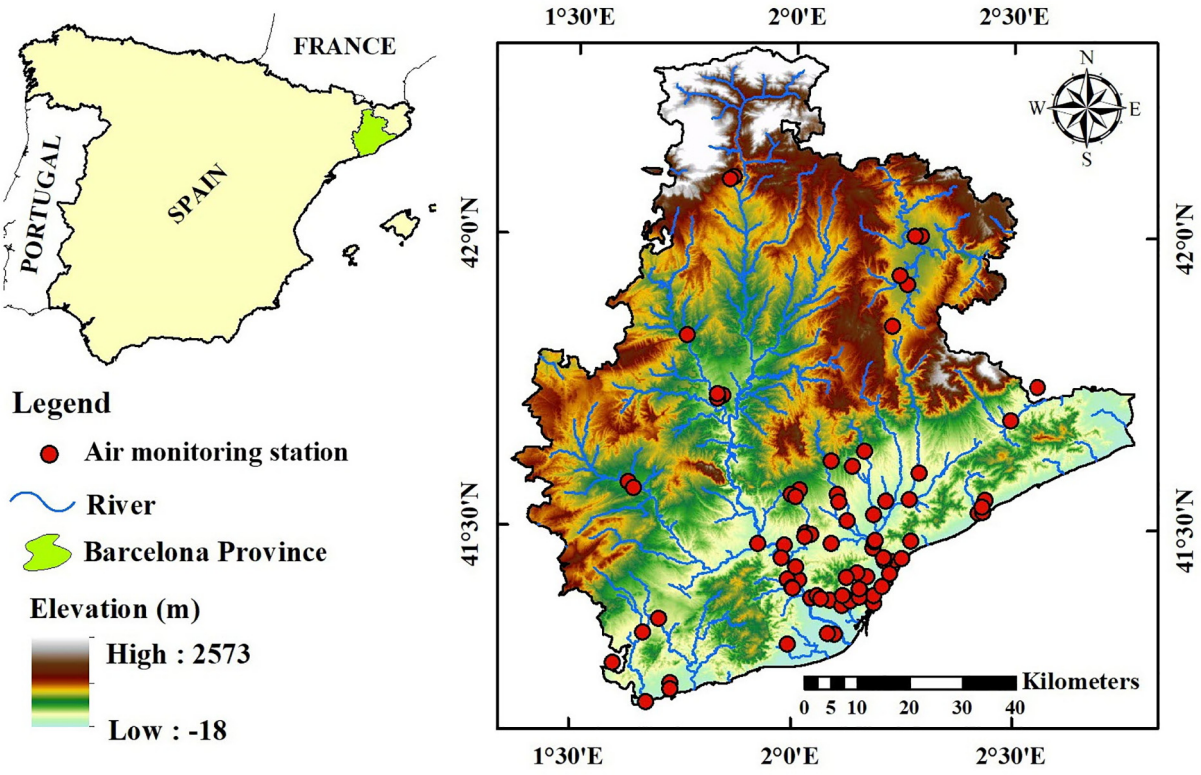


Fig. 1. Location of the Barcelona Province in NE Spain.

to 42°20′ N and longitudes of 5°4′ to 6°28′ E (Fig. 1). The geographic location of this Province is in North-East of Spain with 100 km coastline to the Mediterranean Sea. The Province has about 5,613,955 human population and 7731 km^2 area (Bech et al., 2011; Statistical Institute of Catalonia, 2019). The Provinces surrounding that of Barcelona are those of Tarragona, Girona, Lleida, respectively located in southwest, northeast, northwest of Barcelona. Annual precipitation is around 600 (Bech et al., 2011). The study area has a Mediterranean climate (Clavero et al., 1997). In the west part of Europe, Barcelona is one of the urban and industrial areas suffering severe PM pollution, due to both industrial and road traffic emissions (Querol et al., 2004; Perez et al., 2009, among others), thus predicting air pollution such as PM hazard in this area is very relevant.

2.2. Data used

2.2.1. PM10 ground measurements

The PM data can be divided into three categories: coarse (PM10-2.5, particles with a diameter between 2.5 and 10 μm), intermodal (PM2.5–1, particles with a diameter between 1 and 2.5 μm), and submicron particles (PM1, particles with a diameter less than 1 μm) PM. According to Perez et al. (2009), the cerebrovascular and cardiovascular mortalities in Barcelona are mostly related to PM10-2.5 and PM1 concentrations. They recommended considering these PM data in air pollution mitigation strategies in urban areas. Therefore, due to data availability, in this study, only PM10 was used to hazard assessment of the Barcelona Province. The daily time-series data of PM10 from the Catalonian network of air quality monitoring stations were obtained from the Department of Territory and Sustainability of the Catalonia Government (http://mediambient.gencat.cat/). Some of these were obtained by means of real-time measurements (mostly Beta Attenuation instruments) and in other cases with offline gravimetry, using high volume samplers. The location of the air quality monitoring stations monitoring PM10 is shown in Fig. 1. Due to the fact that the stations have not been established in the same year, there are not long data series in all of them. Therefore, due to data availability, the data were analyzed based on the annual average (μg/m 3) of the years 2015 to 2017 for 75 stations.

After calculating the annual (2015 to 2017) of PM10 averages for each station, all data were surveyed using the healthy threshold presented in European regulations (European Air Quality Framework Directive 2008/50/CE, European Commission, 2008). According to these regulations, a limit value for the protection of human health for yearly PM10 concentrations is 40 $\mu g/m^3$, so based on this threshold the mean yearly concentrations of the PM10 data (from 2015 to 2017) in each station were converted to 0 and 1 respectively indicating the healthy and unhealthy conditions. Eventually, a ratio 70/30 (53/22 stations) was used for calibrating and validating the ML models to predict hazardous areas of PM10 in the Barcelona Province.

2.2.2. Predictors used for PM10 modeling

There are no universal guidelines to select predictors in PM modeling. Here we tried to consider the related variables to PM according to surveys in literature. Accordingly, PM10 as a dependent variable is affected by variables such as land use, soil, meteorological and topographic variables, which are commonly used for modeling air pollution (Li et al., 2017; Kleine Deters et al., 2017; Martínez et al., 2018; Chen et al., 2018b; Choi et al., 2018). Wind data (especially direction and speed) and topography are the essential variables affecting air pollution spatial variability (Kim et al., 2015; Stull, 2012). Lakes and water bodies by changing wind flow and humidity influence also air pollution of a region. (Arain et al., 2007; Kim et al., 2015). Different land uses such as natural areas, agricultural, urban or industrial land use, and roadway as transportation paths are also essential factors (Cohen et al., 2005).

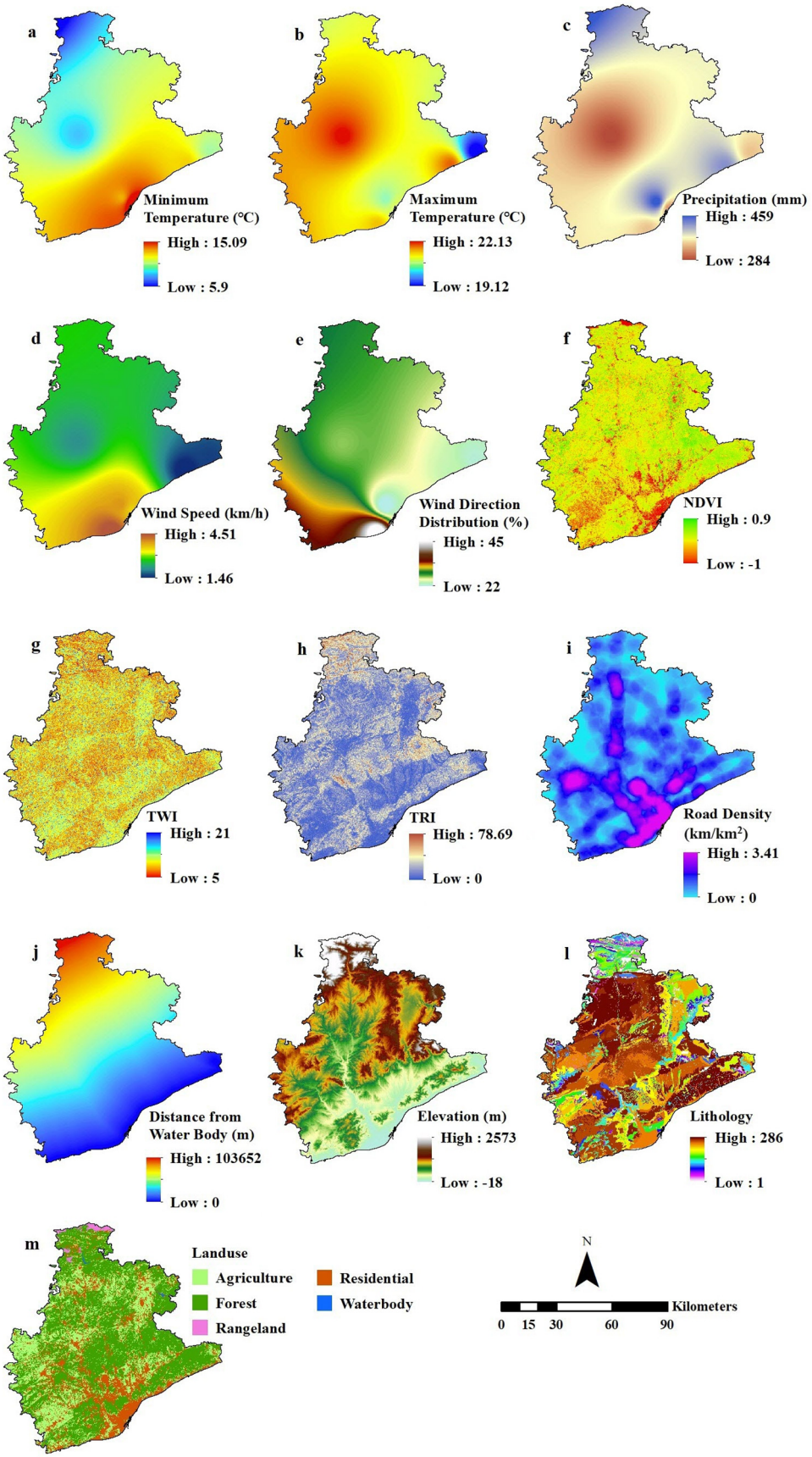


Fig. 2. PM10 conditioning factors: a) minimum temperature, b) maximum temperature, c) precipitation, d) wind speed, e) wind direction distribution, f) normalized difference vegetation index (NDVI), g) topographic wetness index (TWI), h) terrain roughness index (TRI), i) road density, j) distance from water body, k) elevation, l) lithology, m) landuse.

Therefore, 13 important variables (Fig. 2) which were available in the study region, including minimum temperature, maximum temperature, normalized difference vegetation index (NDVI), precipitation, wind speed, wind direction, elevation, road density, topographic wetness index (TWI), land use, terrain roughness index (TRI), distance from water body, and lithology were selected to predict the PM10.

Meteorological data in this study were downloaded from the State Meteorological Agency of Spain (AEMET, http://www.aemet.es/es/) and lithology map was obtained from the Cartographic and Geological Institute of Catalonia (ICGC, http://www.icgc.cat/en/). An ASTER Digital Elevation Model (DEM) 30 m × 30 m was used to extract layers of elevation, TWI, TRI. Maps of land use and NDVI were prepared from the Landsat 8 images for 2017. The layers of road density and distance from water body were prepared using roads and water body layers respectively through Line Density and Euclidean Distance tools in ArcGIS software. PM10 conditioning factors used to prepare the PM10 hazard map have been shown in Fig. 2.

2.3. Feature selection

In machine learning (ML) modeling, where a dependent variable is simulated using the predictor variables, selecting the appropriate numbers of the predictor variables is essential. This is done for the sake of modeling simplification, having more natural interpretation by users, and minimizing the computational cost (James et al., 2013). In the pre-processing of the machine learning, there is a process called Feature Selection (FS), which solves this problem. In this process, redundant variables are identified, and some variables are selected as the most effective predictors (Guyon and Elisseeff, 2003; Jia et al., 2016; Sayed et al., 2019). In this process, the main challenge is to determine the most relevant variables between a dataset.

FS methods are divided into three groups including complete/exhaustive search, heuristics, and non-deterministic search methods (Motoda and Liu, 2002). The complete/exhaustive search and heuristics methods have some problems in finding the best features, such as large search space and reflect a large feature subset (Liu and Motoda, 1998). However, non-deterministic search methods such as simulated annealing (SA) method can address these weaknesses with finding the best solutions through a stochastic search and keeping the one or more of the best solutions which named as elitism (Rudolph, 1994; Meiri and Zahavi, 2006). Therefore, in this study, the SA method was applied for FS selection. Kirkpatrick et al. (1983) first proposed the SA technique according to annealing and cooling of a substance for arriving a low-temperature state with minimal defects in the crystal and minimizing the matter-energy (Rere et al., 2015). In ML modeling, the SA algorithm as a probabilistic technique uses an objective function for combinatorial optimization problems instead of the temperature in a substance. This algorithm for finding an optimum solution among many possible solutions moves from one solution to the next solution. So, each step has a specific possibility for improvement. As the algorithm running the worse solutions are eliminated and arrives in only a step that improvements are accepted (Pinedo, 2016). Here it was necessary to select the best predictor variables among 13 available variables for PM10 prediction. Therefore, this algorithm was used for the FS using the Caret package (Kuhn, 2008; Kuhn, 2015) in R software with the 10-fold cross-validation (CV) method.

2.4. PM 10 hazard modeling

Since the relationship between air pollution and its related variables is complicated, in order to discover these relationships, it is necessary to employ methods that can properly identify and explain them (Bai et al., 2018). Learning machines are one of the approaches used by artificial intelligence to identify complex equations between the relevant variables of air pollution (Delavar et al., 2019). Here, for modeling PM10 hazard, the stations were divided into two datasets for testing and training the ML models including the Mixture Discriminant Analysis (MDA), Bagged Classification and Regression Trees (Bagged CART), and Random Forest (RF). According to scholars in the ML modeling (e.g., Sajedi-Hosseini et al., 2018; Darabi et al., 2019), 70% of the data was used to train the models, and 30% of data was applied for the testing of prediction accuracy. In this study, calibration and validation of the models were done using the Caret package (Kuhn, 2008; Kuhn, 2015) in R software. The leave-one-out CV (LOOCV) and *k*-fold CV methods are the most popular methods which are used to tune the parameters and internal validation (i.e., CV) of the model (Lualdi and Fasano, 2019). These CV methods use different samples (*k* subsets) for model construction. In LOOCV, the number of *k* subsets is equal to the number of samples in the dataset (*n*), this means that the training is done with all dataset and is tested with one sample (leave-one-out), at each round of validation. So, the process is performed *n* times and, eventually, an average error is reported as performance of the model. In *k*-fold CV, the dataset is split into *k* equal size subsets. At each round of validation, the training is conducted using ($k - 1$) subsets and is tested with one of the reminded subsets. The training process is performed *k* times and, an average error is used to indicate the model performance (Lualdi and Fasano, 2019). Since the number of *k* subsets in the LOOCV method is high (i.e., equal to the number of samples in the dataset), it is computationally expensive for some models. So, the LOOCV method usually used to model calibration when the number of datasets is very small. Therefore, in this study, we have applied the *k*-fold cross-validation methodology (k = 10) for calibrating the models using the Caret R package (Kuhn, 2008; Kuhn, 2015). Also, the optimal value of the modeling parameters was tuned using the tuning function of Caret R package. The tuneLength function with a random search which generates the maximum number of tuning parameter combinations was used in this study (Kuhn, 2015). Description of the models is presented as follow:

2.4.1. Mda

MDA is a statistical technique which was first presented by Hair et al. (1998). This algorithm was further developed by Johnson and Wichern (2002) as a data classification approach. The discriminant analysis presents a linear equation as the discriminant function, which consists of independent variables to split the dependent variable. Each independent variable has a weight in the linear equation as discriminant coefficients correct them for finding the best discriminant function. The discriminant equation is the following:

$$F = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon \quad (1)$$

where F is the discriminant score, X_n ($n = 1, 2, 3, \ldots, p$) are independent variables, β_n ($n = 1, 2, 3, \ldots, p$) are the discriminant weights, and ε is the error term (Hair et al., 1998).

2.4.2. Bagged CART

CART (Classification and Regression Trees) is a non-parametric method of decision tree loggings that has been used intensively as a classifier. This method was introduced by Breiman et al. (1984). Its function is based on binary trees. This tree and other trees are the basis for more complex algorithms such as Random Forest. The CART decision tree algorithm divides the data into binary sections to construct the decision tree. The CART tree uses the Gini index to determine which variables gain more information for

classifying data. Variables with lower Gini indexes receive more weight for classification. CART uses the test and error to determine the optimal value for the separator point in each dimension or variable, which has a lower Gini index (Timofeev, 2004).

In the decision trees like CART causing a small change in data may create different classes (these algorithms are also called non-stable algorithms). To solve this problem, the ensemble classifiers are used, which include several classifiers. One of the ensemble classification methods is Bagging. Bagging is among the most popular ensemble techniques for improving the predictive performance of decision trees (Quinlan, 1996). Here, the bagging has been used to ensure the high quality of CART model (Sutton, 2004). In this method, each classifier builds its model by classification of the part of data and stores it. Finally, using vote intention among these classifications, the class that receives the most votes is used as the final classifier (Quinlan, 1996; Breiman, 1996).

2.4.3. Rf

RF is a combinatorial algorithm that uses Decision Trees for data classification (Breiman, 2001). In the combinatorial algorithms, several classification algorithms for decision making and splitting each node are used. Each node is a homogeneous region, including observational data, which are categorized and split by algorithms. The combinatorial classification generally prevents the overfitting of the model learned by the algorithm and in many cases, produces better results than other algorithms (Cutler et al., 2007). The RF model grants a subset of the data to each decision tree, which is randomly selected. Then each decision tree can decide according to the individual data subset and can make their classification. To classify new data, according to the prediction result of each of the algorithms, the RF selects just one of them as the final algorithm for classification. Each variable has a weight value that is used in the classification and varies randomly according to misclassification for observational data to evaluate the importance of each predictor. Finally, the model is used for the new prediction when the lowest misclassification occurred. The RF using this method improves classification accuracy (Breiman, 2001).

2.5. Model evaluation

One of the most important steps after modeling is to evaluate the performance of the model. Model evaluation helps to modify the structure of the model in applying variables for proper classification. The hit and miss analysis through a confusion matrix are used to evaluate the performance of the models for dichotomous (yes/no) forecasts. So, in this study according to literature (Johnson and Olsen, 1998; Sokolova et al., 2006; Choubin et al., 2019a), the models' evaluation was performed using the five statistics including Accuracy, Precision, Bias, Probability of Detection (POD) and False Alarm Ratio (FAR) (Eq (2) to (6)). These indices are calculated using the hit and miss analysis through a confusion matrix. The confusion matrix is a record of possible classifications of correctly and incorrectly classified data. All indices are between 0 and 1, which 0 in FAR indicates the best classification while for others 1 indicates the best performance (Johnson and Olsen, 1998; Sokolova et al., 2006; Choubin et al., 2019a). Equation of the metrics are presented as follows:

$$\text{Accuracy} = \frac{H + CN}{H + FA + M + CN} \quad (2)$$

$$\text{Precision} = \frac{H}{H + FA} \quad (3)$$

$$\text{Bias} = \frac{H+FA}{H+M} \quad (4)$$

$$\text{POD} = \frac{H}{H+M} \quad (5)$$

$$\text{FAR} = \frac{FA}{H+FA} \quad (6)$$

where H is Hits, FA is False Alarms, M is Misses, and CN is Correct Negative in a confusion matrix (Sokolova et al., 2006).

3. Results and discussion

3.1. Feature selection results

In this study, the key features were selected using the SA method described above. According to the SA results, a number of the selected features varied between 2 and 7 in each fold based on the objective function accuracy (Table 1). For instance, in the fold01, the 4 features (i.e., DFWB, Landuse, NDVI, and TWI) indicated that a higher accuracy (equal to 80%) in comparison to other possible combinations. In Table 1, critical features selected by the SA in each fold are represented. To select critical features for the whole dataset (all folds), we investigated the frequency (%) of the selected features according to the all fold's information (Fig. 3). Results indicated that the key features are respectively TWI, land use, road density, precipitation, distance from the water body, NDVI, TRI, maximum temperature, minimum temperature, wind speed, and elevation with a frequency of 90%, 70%, 50%, 50%, 50%, 40%, 40%, 30%, 20%, 10%, and 10%, respectively. The lithology and wind direction were not selected in any fold as key features (Fig. 3). According to the SA results, although we could consider the number of 2 to 7 variables as key features, we considered maximum possible features (i.e., 7 features due to the maximum number of the selected features in the folds, Table 1) according to their occurrence frequency in the 10-fold. Therefore, variables of TWI, land use, road density, precipitation, distance from water body, and NDVI which had higher frequencies (as selected features in the folds) were selected as key features for PM10 modeling. These features had >40% frequency in the folds (Fig. 3).

3.2. Evaluation of the model's performance

Performance of the models in the prediction of PM10 is presented in Table 2. According to the results, the models had high accuracy (0.87, 0.92, and 0.93, respectively, for the MDA, Bagged CART, and RF models) (Table 2). Given the precision, the Bagged CART and RF models had a close precision (0.86 and 0.87 respectively), which were higher than the MDA (0.69) model. The MDA model had a perfect Bias means that the summation of the false alarms (FA) and misses (M) was equal, so the model had not an

Table 1
Key features based on the accuracy after 100-iteration in each fold using the simulated annealing method.

Fold	Number of the selected features	Selected features	Accuracy
Fold01	4	DFWB, Landuse, NDVI, TWI	0.80
Fold02	4	Landuse, NDVI, Tmax, TRI	0.75
Fold03	4	DFWB, Landuse, NDVI, TWI	0.82
Fold04	5	DFWB, Landuse, PCP, Tmin, TWI	0.82
Fold05	2	DFWB, TWI	0.78
Fold06	5	Landuse, PCP, RD, TRI, TWI	0.80
Fold07	4	Landuse, RD, TRI, TWI	0.83
Fold08	6	Landuse, PCP, RD, Tmin, TWI, WS	0.81
Fold09	5	NDVI, PCP, RD, Tmax, TWI	0.79
Fold10	7	DFWB, Elevation, PCP, RD, Tmax, TRI, TWI	0.81

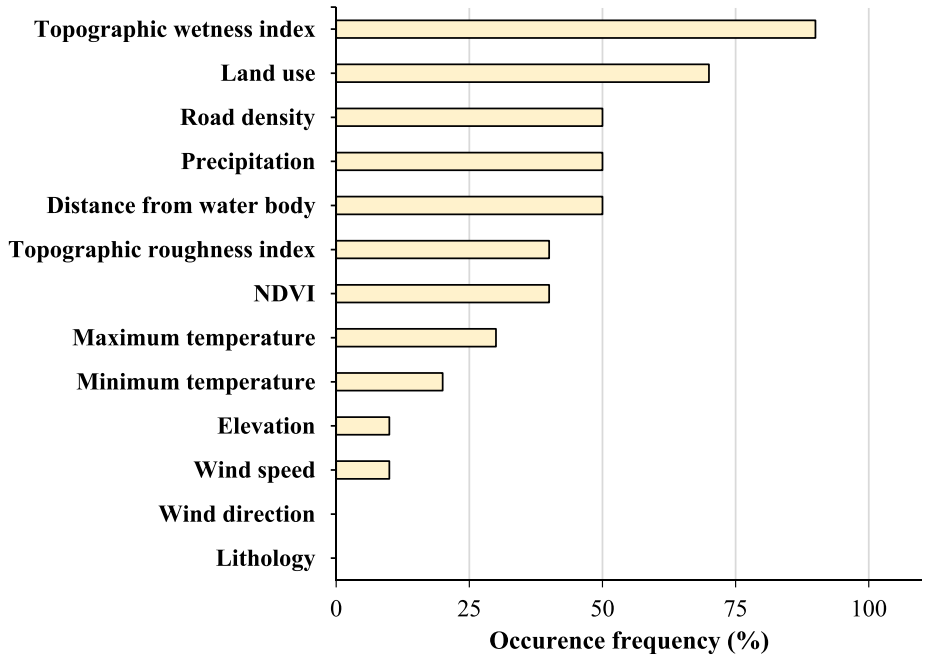


Fig. 3. The frequency (%) of the selected features by simulated annealing method in 10-fold.

Table 2
Performance of the models in prediction of the PM10.

Statistic	MDA	Bagged CART	RF
Accuracy	0.87	0.92	0.93
Precision	0.69	0.86	0.87
Bias	1.00	0.88	0.94
POD	0.69	0.75	0.81
FAR	0.31	0.14	0.13

overestimate or underestimate. Whereas both of the Bagged CART and RF models indicated an underestimate (respectively Bias was equal to 0.88 and 0.94). The probability of detection (POD) was higher critical respectively for the RF, Bagged CART, and MDA models (respectively 0.81, 0.75, and 0.69). Also, false alarm ratio (FAR) in the Bagged CART and RF models were close precision (0.14 and 0.13 respectively) and lower than the MDA (0.31) model (Table 2).

In PM10 modeling, although the application of the MDA and Bagged CART models were novel, generally, all the three-machine learning (ML) models indicated an excellent performance (Accuracy > 85%). However, the Bagged CART and RF models had the same performance and higher than the MDA model. Former studies, e.g., Chen et al. (2018a) showed that the RF model has a higher predictive accuracy than traditional regression models to predict the PM2.5 in China. In another study presented by Martínez et al. (2018), the RF model outperformed the conventional CART and logistic regression models. Thus, the ensemble techniques generally appear to improve the performance of the models. Consequently, it is expected that novel ensemble ML models contribute further in the advancement of the future's high-performance models, as it has been the case in the hydrological and the other earth systems hazard assessment models (Yaseen et al. 2019; Fotovatikhah et al. 2018; Moazenzadeh et al. 2018; Mosavi et al. 2018). The future research on spatial hazard assessment of the various PM can also benefit from hybrid ML models. The performance of ML methods has showed to be dramatically improved through hybridization with statistical methods, soft computing techniques and optimization algorithms. However, the application of hybrid ML methods in air pollution and PM hazard assessment is in the early stage. Future research can highly benefit from numerous advantages of hybrid models.

3.3. Spatial hazard assessment

After validation of the models, the pixels' value of the predictive variables was used to predict the probability of the healthy and unhealthy areas in view of PM10 for the whole study. Then the spatial hazard maps were produced by classifying the probability maps into low, medium, and high classes by natural break classification in the ArcGIS environment (Fig. 4). On the MDA model map, classes of low, high, and medium hazards covered the highest towards the lowest area, respectively, equal to 3162.0 km^2 (40.9%), 2506.5 km^2 (32.4%), and 2062.2 km^2 (26.7%). On the Bagged CART model map, the highest area (3648.2 km^2; 47.2%) is related to the low class, while the high class has the lowest area (1013.9 km^2; 13.1%). In this model, the medium class has covered about 39.7% (3068.6 km^2) of the study area. According to the RF model map, the medium class has the highest area (3989.5 km^2; 51.6%), and the high class (2255.6 km^2; 29.2%) has a higher area than low class (1485.6 km^2; 19.2%) (Table 3).

The hazard classes produced by the three models do not entirely match. The difference comes from the different structure of the models. However, areas in the middle of Barcelona Province indicated a high hazard by the three models (Fig. 4). Contribution of the predictive variables for the best model (i.e., the RF model) is represented in Fig. 5. As can be seen, variables of topographic wetness index (TWI), and topographic roughness index (TRI) have the highest contribution to PM10 hazard modeling by the RF model. Also, land use has the lowest contribution in predicting the PM10 hazard map (Fig. 5).

According to the results hazardous zone are located in the mid less industrialized and urbanized areas of the Province than in the Barcelona's Metropolitan Area. This unexpected pattern might be due to i) the frequent inland transport of pollution, ii) the more abrupt topography of the mid areas, iii) the high ammonia (NH_3) levels recorded in the mid areas of the province (Van Damme et al., 2018), and iv) to the increased emissions of domestic and commercial biomass burning in small cities and villages. In fact,

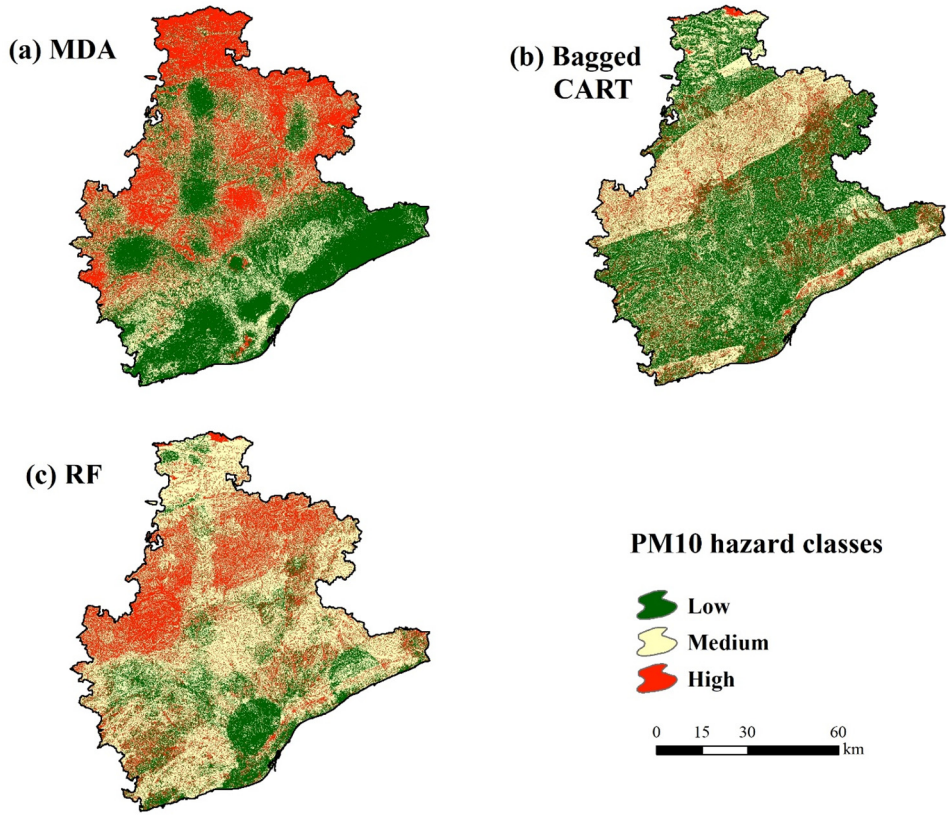


Fig. 4. PM10 hazard maps produced by MDA (a), Bagged CART (b), and RF (c) models.

Table 3
The area of the classes of PM10 hazard.

Class	Area	MDA	Bagged CART	RF
Low	km^2	3162.0	3648.2	1485.6
	%	40.9	47.2	19.2
Medium	km^2	2062.2	3068.6	3989.5
	%	26.7	39.7	51.6
High	km^2	2506.5	1013.9	2255.6
	%	32.4	13.1	29.2

one of the sites recording the exceedances of the daily limit value in the last years in the mid areas of the Province was also recording exceedances of the benzo[a]pyrene target value of the above air quality directive, and this was attributed to biomass burning emissions. The transport and chemical PM10 modelling outputs tend to yield higher PM10 levels in the Barcelona's Metropolitan Area, although the PM10 pollution maps tend to vary from one year to

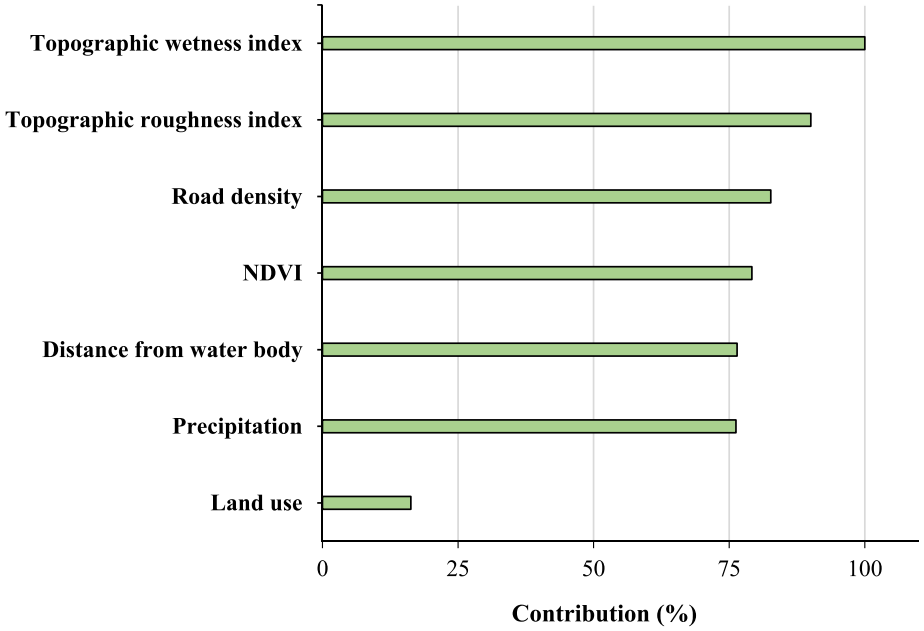


Fig. 5. Contribution of the predictive variables in the RF model.

the other according meteorology and emission patterns (Generalitat de Catalunya 2015; 2016, and BSC-CALIOPE et al., 2013). Some of the possible reasons might be i) the more abrupt topography of the mid and north regions that favors the accumulation of pollutants in valleys and intra-mountain plains, ii) a possible sub-estimation of the emission inventories of the domestic and commercial biomass burning, and iii) the inaccurate NH_3 emission inventories from farming and agriculture. The latter being a key gaseous pollutant to generate secondary PM, and Catalonia being a European hotspot of NH_3, according results from remote sensing (Van Damme et al., 2018). In deep research is needed to find the actual causes for this mismatching. In any case, PM10 recorded across the region are exceeding the respective WHO air quality guideline, and accordingly efforts shall be done to abate emissions of PM10 and gaseous precursors across the study area.

The method of spatial hazard prediction of the PM10 proposed in this paper can be used in hazard prediction of other air quality parameters. Furthermore, the hazard mapping in respect to the air quality index (AQI) can be predicted using the proposed method by considering one or more air pollutants, e.g., nitrogen dioxide (NO_2), carbon monoxide (CO), ground-level ozone (O_3), ammonia (NH_3), and volatile organic compounds (VOC), among others.

4. Conclusion

The current research tried to predict the spatial hazard of the PM10 using the machine learning models (i.e., Bagged Classification and Regression Trees, CART; Mixture Discriminate Analysis, MDA; and Random Forest, RF). The number of the 13 effective factors were considered and reduced by the simulated annealing (SA) feature selection method. The SA results indicated that the variables of topographic wetness index (TWI), land use, road density, precipitation, distance from a water body, and normalized difference vegetation index (NDVI) were the critical features for PM10 modeling. According to the results, the models had high accuracy (precision) equal to 0.87 (0.69), 0.92 (0.86), and 0.93 (0.87) respectively for the MDA, Bagged CART, and RF models. Thus, the Bagged CART and RF models had the same performance and higher than the MDA model. Furthermore, the modeling results by the three models indicated that areas in the middle of Barcelona Province had a higher hazard than the more industrialized and urbanized areas of the Barcelona's Metropolitan Area. Also, the variables of TWI, and topographic roughness index (TRI) had the highest contribution to PM10 hazard modeling, while the land use had the lowest contribution. Some of the possible reasons for the higher PM10 hazard found for the mid areas of the Province might include the following ones: i) the more abrupt topography of the mid and north regions that favors the accumulation of pollutants in valleys and intra-mountain plains, ii) the weakness of high emission from residential and agricultural biomass burning, and iii) high NH_3 emission from farming and agriculture. However, in deep research is needed to find the actual causes for this mismatching. Although our findings highlighted a good hazard prediction of the PM10 concerning the environmental conditioning factors, there were some inevitable limitations. In this study, the PM10 data for the simultaneous measurement in the stations was short, so considering a longer period of data caused a decrease in the number of monitoring stations which is most important in the ML modeling. Hence, due to data availability and ML modeling requirement, the possible maximum number of stations from 2015 to 2017 was used to produce the hazard maps. Providing a longer period of PM data can help in producing more reliable hazard maps. Another inevitable limitation is related to the source of PM data, as some sources of the PM10 may be from outside of the region and identifying and considering this in the modeling process is too difficult. However, the proposed models presented a high accuracy in prediction of hazardous areas of PM10. Sustainable urban development, environmental management, and land-use policies highly depend on accurate hazard maps for policy-making. Therefore, the important contributors in increasing the PM10 can better come to consideration for mitigation and resilience policies.

Acknowledgments

The authors would like to thank the Department of Territory and Sustainability, and the Institute of Cartography and Geology of Catalonia, both from the Generalitat de Catalunya, and the State Meteorological Agency, from the Ministry for the Ecological Transition of Spain, for supplying us with the PM data, lithology maps and meteorological data, respectively.

References

Abbey, D., Nishino, N., McDonnell, W., Burchette, R., Knustsen, S., Beeson, W., Yang, J., 1999. Long-term inhalable particles and other air pollutants related to mortality in non-smokers. Am. J. Respir. Crit. Care Med. 159, 373–382.

Amodio, M., Bruno, P., Caselli, M., de Gennaro, G., Dambruoso, P.R., Daresta, B.E., et al., 2008. Chemical characterization of fine particulate matter during peak PM10 episodes in Apulia (South Italy). Atmos. Res. 90, 313–325.

Arain, M.A., Blair, R., Finkelstein, N., Brook, J.R., Sahsuvaroglu, T., Beckerman, B., Jerrett, M., 2007. The use of wind fields in a land use regression model to predict air pollution concentrations for health exposure studies. Atmos. Environ. 41 (16), 3453–3464.

Bai, L., Wang, J., Ma, X., Lu, H., 2018. Air pollution forecasts: an overview. Int. J. Environ. Res. Public Health 15 (4), 780.

Bech, J., Tume, P., Sánchez, P., Reverter, F., Bech, J., Lansac, A., Oliver, T., 2011. Levels and pedogeochemical mapping of lead and chromium in soils of Barcelona province (NE Spain). J. Geochem. Explor. 109 (1–3), 104–112.

Beelen, R., Raaschou-Nielsen, O., Stafoggia, M., et al., 2014. Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 European cohorts within the multicentre ESCAPE project. Lancet 383, 785–795.

Breiman, L., 1996. Bagging predictors. Machine learning 24 (2), 123–140.

Breiman, L., 2001. Random forests. Machine learning 45 (1), 5–32.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and regression trees, vol. 432. Wadsworth International Group, Belmont, CA, pp. 151–166.

Brook, R.D., Rajagopalan, S., Pope 3rd, C.A., et al., 2010. American Heart Association Council on Epidemiology and Prevention, Council on the Kidney in Cardiovascular Disease, and Council on Nutrition, Physical Activity and Metabolism. Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the American Heart Association. Circulation 121, 2331–2378.

Brunekreef, B., Holgate, S.T., 2002. Air pollution and health. Lancet 360, 1233e1242.

Brunelli, U., Piazza, V., Pignato, L. and Sorbello, F.and Vitabile, S. (2007). Two-day ahead prediction of daily maximum.

BSC-CALIOPE, 2013. Map of PM10 levels over Spain in 2013. Report on the evaluation of the 2013 CALIOPE Forecast System. http://www.bsc.es/projects/earthscience/visor/bases_datos/image_viewer/docs/20140513_Informe_Evaluacion_Pronostico_CALIOPE_2013.pdf

Chen, B., Kan, H., 2008. Air pollution and population health: a global challenge. Environ Health Prev Med 13 (2), 94–101.

Chen, G., Li, S., Knibbs, L.D., Hamm, N.A.S., Cao, W., Li, T., Guo, J., Ren, H., Abramson, M.J., Guo, Y., 2018a. A machine learning method to estimate PM 2.5 concentrations across China with remote sensing, meteorological and land use information. Sci. Total Environ. 636, 52–60.

Chen, G., Wang, Y., Li, S., Cao, W., Ren, H., Knibbs, L.D., Abramson, M.J., Guo, Y., 2018b. Spatiotemporal patterns of PM10 contaminations over China during 2005_2016: a satellite-based estimation using the random forests approach. Environ. Pollut. 242, 605e613.

Chen, M.J., Yang, P.H., Hsieh, M.T., Yeh, C.H., Huang, C.H., Yang, C.M., Lin, G.M., 2018. Machine learning to relate PM2.5 and PM10 contaminations to outpatient visits for upper respiratory tract infections in Taiwan: a nationwide analysis. World J. Clin. Cases 6(8): 200-206.

Chen, Y.Y., Shi, R.H., Shu, S.J., Gao, W., 2013. Ensemble and enhanced PM10 contamination forecast model based on stepwise regression and wavelet analysis. Atmos. Environ. 74, 346–359.

Choi, J.E., Lee, H., Song, J., 2018. Forecasting daily PM10 concentrations in Seoul using various data mining techniques. Communications for Statistical Applications and Methods 25 (2), 199–215.

Choubin, B., Borji, M., Mosavi, A., Sajedi-Hosseini, F., Singh, V.P., Shamshirband, S., 2019a. Snow avalanche hazard prediction using machine learning methods. J. Hydrol. 123929

Choubin, B., Moradi, E., Golshan, M., Adamowski, J., Sajedi-Hosseini, F., Mosavi, A., 2019b. An Ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines. Sci. Total Environ. 651, 2087–2096.

Clavero, P., Martin Vide, J., Raso, J.M., 1997. Atles Climàtic deCatalunya. Barcelona, I. C.C. Dept. de Medi Ambient 42, maps.

Cobourn, W.G., 2010. An enhanced PM2.5 air quality forecast model based on nonlinear regression and back-trajectory contaminations. Atmos. Environ. 44 (25), 3015–3023.

Cohen, J., Cook, R., Bailey, C.R., Carr, E., 2005. Relationship between motor vehicle emissions of hazardous pollutants, roadway proximity, and ambient concentrations in Portland. Oregon. Environmental Modelling & Software 20 (1), 7–12.

Corani, G., 2005. Air quality prediction in Milan: Feed-forward neural networks, pruned neural networks and lazy learning. Ecol Model 185 (2–4), 513–529.

Cutler, D.R., Edwards Jr, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. Ecology 88 (11), 2783–2792.

Dall'Osto, M., Beddows, D.C.S., Pey, J., Rodriguez, S., Alastuey, A., Harrison, Roy M., Querol, X., 2012. Urban aerosol size distributions over the Mediterranean city of Barcelona, NE Spain. Atmos. Chem. Phys. 12, 10693–10707.

Daniel, J., Jacob., Darrell, A, Winner.,, 2009. Effect of climate change on air quality. Atmos. Environ. 43, 51–63.

Darabi, H., Choubin, B., Rahmati, O., Haghighi, A.T., Pradhan, B., Kløve, B., 2019. Urban flood risk mapping using the GARP and QUEST models: A comparative study of machine learning techniques. J. Hydrol. 569, 142–154.

De Gennaro, G., Trizio, L., Di Gilio, A., Pey, J., Pérez, N., Cusack, M., Alastuey, A., Querol, X., 2013. Neural network model for the prediction of PM10 daily contaminations in two sites in the Western Mediterranean. Sci. Total Environ. 463–464, 875–883.

Dehghan, A., Khanjani, N., Bahrampour, A., Goudarzi, G., Yunesian, M., 2018. The relation between air pollution and respiratory deaths in Tehran, Iran-using generalized additive models. BMC Pulm. Med. 18, 49.

Delavar, M.R., Gholami, A., Shiran, G.R., Rashidi, Y., Nakhaeizadeh, G.R., Fedra, K., Hatefi Afshar, S., 2019. A Novel Method for Improving Air Pollution Prediction Based on Machine Learning Approaches: A Case Study Applied to the Capital City of Tehran. ISPRS Int. J. Geo-Inf. 8 (2), 99.

Djalalova, I., Monache, L.D., Wilczak, J., 2015. PM2.5 analog forecast and Kalman filter post-processing for the Community Multi-scale Air Quality (CMAQ) model. Atmos. Environ. 108, 76–87.

Dominici, F., Peng, R.D., Bell, M.L., Pham, L., Mcdermott, A., Zeger, S.L., Samet, J.M., 2006. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. JAMA 295, 1127–1134.

Donaldson, K., MacNee, W., 1999. The mechanism of lung injury caused by PM10. In: Air Pollution and Health. In: Hester, R.E., Harrison, R.M. (Eds.), Issue in Environmental Science and Technology. Royal Society of Chemistry, Reedwood Books Ltd., Trowbridge, Wiltshire, UK.

Elangasinghe, M.A., Singhal, N., Dirks, K.N., Salmond, J.A., 2014. Development of an ANN-based air pollution forecasting system with explicit knowledge through sensitivity analysis. Atmos. Pollut. Res., 5

Escudero, M., Querol, X., Ávila, A., Cuevas, E., 2007. Origin of the exceedances of the European daily PM limit value in regional background areas of Spain. Atmos. Environ. 41 (4), 730–744.

European Commission, P., 2008. Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. Official Journal of the European Communities, L 152 (2008), 1–44.

Feng, W., Li, H., Wang, S., Van Halm-Lutterodt, N., An, J., Liu, Y., Guo, X., 2019. Short-term PM10 and emergency department admissions for selective cardiovascular and respiratory diseases in Beijing, China. Sci. Total Environ. 657, 213–221.

Feng, X., Li, Q., Zhu, Y.J., Hou, J.X., Jin, L.Y., Wang, J.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.

Fernández-Navarro, P., Garcıa-Pérez, J., Ramis, R., Boldo, E., López-Abente, G., 2017. Industrial pollution and cancer in Spain: an important public health issue. Environ. Res. 159, 555–563.

Fortelli, A., Scafetta, N., Mazzarella, A., 2016. Influence of synoptic and local atmospheric patterns on PM10 air pollution levels: a model application to Naples (Italy). Atmos. Environ. 143, 218–228.

Fotovatikhah, F., Herrera, M., Shamshirband, S., Chau, K.W., Faizollahzadeh Ardabili, S., Piran, M.J., 2018. Survey of computational intelligence as basis to big flood management: challenges, research directions and future work. Engineering Applications of Computational Fluid Mechanics 12 (1), 411–437.

Franklin, M., Koutrakis, P., Schwartz, J., 2008. The role of particle composition on the association between $PM_{2.5}$and mortality. Epidemiology 19, 680–689.

Ganguly, R., Sharma, D., Kumar, P., 2019. Trend analysis of observational PM10 concentrations in Shimla city, India. Sustainable Cities and Society 51, 101719.

García Nieto, P.J., Sánchez Lasheras, F., García-Gonzalo, E., 2018. Estimation of PM10 contamination from air quality data in the vicinity of a major steelworks site in the metropolitan area of Avilés (Northern Spain) using machine learning techniques. Stochastic Environmental Research and Risk Assessment 32 (11), 3287.

Gardner, M.W., Dorling, S.R., 1998. Artificial neural networks: the multilayer perceptron—a review of applications in atmospheric sciences. Atmos. Environ. 32 (14/15), 2627–2636.

Generalitat de Catalunya 2015. Map of ambient PM10 levels over Catalunya. Annual mean 2015. http://mediambient.gencat.cat/web/.content/home/ambits_dactuacio/atmosfera/qualitat_de_laire/avaluacio/analisi_anual/PM10_mean.jpg.

Generalitat de Catalunya 2016. Map of ambient PM10 levels over Catalunya: Annual mean 2016. http://mediambient.gencat.cat/web/.content/home/ambits_dactuacio/atmosfera/qualitat_de_laire/avaluacio/analisi_anual/PM10_mean.jpg.

Gupta, P., Christopher, S.A., 2009. Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: multiple regression approach. J. Geophys. Res. Atmos., p. 114.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Machine Learning Res. 3 (Mar), 1157–1182.

Hair, J.F., Anderson, R.E., Tatham, R.L., Black, W.C., 1998. Multivariate data analysis. Englewood Cliffs, New Jersey, SA.

Hoek, G., Brunekreef, B., Goldbohm, S., Fischer, P., van den Brandt, P.A., 2002a. Association between mortality and indicators of traffic-related air pollution in the Netherlands: a cohort study. The lancet 360 (9341), 1203–1209.

Hoek, G., Brunekreef, B., Goldbohm, S., Fischer, P., Van den Brant, P., 2002. Association between mortality and indicators.

Hrust, L., Klaic, Z.B., Krizan, J., Antonic, O., Hercog, P., 2009. Neural network forecasting of air pollutants hourly contaminations using optimised temporal averages of meteorological variables and pollutant contaminations. Atmos. Environ. 43, 5588–5596.

Hu, X., Belle, J.H., Meng, X., Wildani, A., Waller, L., Strickland, M., Liu, Y., 2017. Estimating PM2. 5 contaminations in the conterminous United States using the randomforest approach. Environ. Sci. Technol. 51, 6936–6944.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Jia, H., Ding, S., Du, M., Xue, Y., 2016. Approximate normalized cuts without Eigen-decomposition. Inf. Sci. 374, 135–150.

Johnson, L.E., Olsen, B.G., 1998. Assessment of quantitative precipitation forecasts. Weather Forecasting 13 (1), 75–83.

Kim, K.H., Lee, S.B., Woo, D., Bae, G.N., 2015. Influence of wind direction and speed on the transport of particle-bound PAHs in a roadway environment. Atmos. Pollut. Res. 6 (6), 1024–1034.

Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., 1983. Optimization by simulated annealing. Science 220 (4598), 671–680.

Kleine Deters, J., Zalakeviciute, R., Gonzalez, M., & Rybarczyk, Y., 2017. Modeling PM2. 5 urban pollution using machine learning and selected meteorological parameters. J. Electr. Comput. Eng. 2017.

Kloog, I., Chudnovsky, A.A., Just, A.C., Nordio, F., Koutrakis, P., Coull, B.A., Lyapustin, A., Wang, Y., Schwartz, J., 2014. A new hybrid spatio-temporal model for estimating daily multi-year PM2.5 contaminations across northeastern USA using high resolution aerosol optical depth data. Atmos. Environ. 95, 581–590.

Konovalov, I.B., Beekmann, M., Meleux, F., Dutot, A., Foret, G., 2009. Combining deterministic and statistical approaches for PM10 forecasting in Europe. Atmos. Environ. 43, 6425–6434.

Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28, 1–26.

Kuhn, M., 2015. Caret: classification and regression training. Astrophysics Source Code Library.

Kumar, U., Ridder, K.D., 2010. GARCH modelling in association with FFT-ARIMA to forecast ozone episodes. Atmos. Environ. 44 (34), 4252–4265.

Lary, D.J., Laery, T., Sattler, B., 2015. Using Machine Learning to Estimate Global PM2.5 for Environmental Health Studies. Environ. Health Insights 9 (S1), 41–52. https://doi.org/10.4137/EHI.S15664.

Lee, H., Liu, Y., Coull, B., Schwartz, J., Koutrakis, P., 2011. A novel calibration approach of MODIS AOD data to predict PM2. 5 contaminations. Atmos. Chem. Phys. 11, 7991.

Lei, M., Luan, S.Y., Jiang, C.W., Liu, H.L., Zhang, Y., 2009. A review on the forecasting of wind speed and generated power. Renew. Sust. Energ. Rev. 13 (4), 915–920.

Li, T., Shen, H., Yuan, Q., Zhang, X., Zhang, L., 2017. Estimating ground-level PM2. 5 by fusing satellite and station observations: a geo-intelligent deep learning approach. Geophys. Res. Lett., 44(23)

Liu, D.J., Li, L., 2015. Application study of comprehensive forecasting model based on entropy weighting method on trend of PM2.5 contamination in Guangzhou, China. Int. J. Environ. Res. Public Health 12, 7085–7099.

Liu, H., Motoda, H. eds., 1998. Feature extraction, construction and selection: a data mining perspective (Vol. 453). Springer Science & Business Media.

Liu, Y., Liu, W., Xu, Y., Zhao, Y., Wang, P., Yu, S., Liu, W., 2019. Characteristics and human inhalation exposure of ionic per-and polyfluoroalkyl substances (PFASs) in PM10 of cities around the Bohai Sea: Diurnal variation and effects of heating activity. Sci. Total Environ. 687, 177–187.

Liu, Y., Paciorek, C.J., Koutrakis, P., 2009. Estimating regional spatial and temporal variability of PM2. 5 contaminations using satellite data, meteorology, and land use information. Environ. Health Perspect. 117, 886.

Lualdi, M., Fasano, M., 2019. Statistical analysis of proteomics data: a review on feature selection. J. Proteomics 198, 18–26.

Luo, H., Wang, D., Yue, C., Liu, Y., Guo., H., 2018. Research and application of a novel hybrid decomposition-ensemble learning paradigm with error correction for daily PM10 forecasting. Atmos. Res. 201, 34–45.

Marchetti, S., Longhin, E., Bengalli, R., Avino, P., Stabile, L., Buonanno, G., Mantecca, P., 2019. In vitro lung toxicity of indoor PM10 from a stove fueled with different biomasses. Sci. Total Environ. 649, 1422–1433.

Martínez, N.M., Montes, L.M., Mura, I. and Franco, J.F., 2018, November. Machine Learning Techniques for PM 10 Levels Forecast in Bogotá. In 2018 ICAI Workshops (ICAIW) (pp. 1-7). IEEE.

Meiri, R., Zahavi, J., 2006. Using simulated annealing to optimize the feature selection problem in marketing applications. Eur. J. Oper. Res. 171 (3), 842–858.

Moazenzadeh, R., Mohammadi, B., Shamshirband, S., Chau, K.W., 2018. Coupling a firefly algorithm with support vector regression to predict evaporation in northern Iran. Engineering Applications of Computational Fluid Mechanics 12 (1), 584–597.

Mosavi, A., Ozturk, P., Chau, K.W., 2018. Flood prediction using machine learning models: literature review. Water 10 (11), 1536.
Motoda, H., Liu, H., 2002. Data reduction: feature selection. In Handbook of data mining and knowledge discovery. Oxford University Press Inc., pp. 208–213.
Nabavi, S.O., Haimberger, L., Abbasi, E., 2019. Assessing PM2. 5 contaminations in Tehran, Iran, from space using MAIAC, deep blue, and dark target AOD and machine learning algorithms. Atmos. Pollution Res. 10 (3), 889–903.
Niu, M.F., Wang, Y.F., Sun, S.L., Li, Y.W., 2016. A novel hybrid decomposition-and-ensemble model based on CEEMD and GWO for short-term PM2.5, contamination forecasting. Atmos. Environ. 134, 168–180.
Oprea, M., Mihalache, S.F., Popescu, M., 2016. A comparative study of computational intelligence techniques applied to PM2.5 air pollution forecasting. In: International Conference on Computers Communications and Control. IEEE, pp. 103–108.
Peng, R., Bell, M., Geyh, A., McDermott, A., Zeger, S., Samet, J., et al., 2009. Emergency admissions for cardiovascular and respiratory diseases and the chemical composition of fine particle air pollution. Environ. Health Perspect. 117, 957–963.
Perez, L., Medina-Ramón, M., Kunzli, N., Alastuey, A., Pey, J., Pérez, N., Sunyer, J., 2009. Size fractionate particulate matter, vehicle traffic, and case-specific daily mortality in Barcelona, Spain. Environ. Sci. Technol. 43 (13), 4707–4714.
Pinedo, M.L., 2016. Scheduling: Theory, Algorithms, and Systems. Springer, New York.
Pope, C., Burnet, R., Thun, M., et al., 2002. Lung cancer, cardiopulmonary mortality, and long term exposure to fine particulate air pollution. JAMA 287, 1132–1141.
Querol, X., Pey, J., Pandolfi, M., Alastuey, A., Cusack, M., Pérez, N., et al., 2009. African dust contributions to mean ambient PM10 mass-levels across the Mediterranean Basin. Atmos. Environ. 43, 4266–4277.
Querol, X., Alastuey, A., Ruiz, C.R., Artiñano, B., Hansson, H.C., Harrison, R.M., Buringh, E.T., Ten Brink, H.M., Lutz, M., Bruckmann, P., Straehl, P., 2004. Speciation and origin of PM10 and PM2. 5 in selected European cities. Atmos. Environ. 38 (38), 6547–6555.
Quinlan, J. R. (1996, August). Bagging, boosting, and C4. 5. In AAAI/IAAI, Vol. 1 (pp. 725-730).
Rere, L.R., Fanany, M.I., Arymurthy, A.M., 2015. Simulated annealing algorithm for deep learning. Procedia Comput. Sci. 72, 137–144.
Rodriguez, S., Querol, X., Alastuey, A., Kallos, G., Kakaliagou, O., 2001. Saharan dust contribution to PM10 and TSP levels in Southern and Eastern Spain. Atmos. Environ. 35, 2433–2447.
Rückerl, R., Schneider, A., Breitner, S., et al., 2011. Health effects of particulate air pollution: a review of epidemiological evidence. Inhal. Toxicol. 23, 555–592.
Rudolph, G., 1994. Convergence analysis of canonical genetic algorithms. IEEE Trans. Neural Networks 5 (1), 96–101.
Sajedi-Hosseini, F., Malekian, A., Choubin, B., Rahmati, O., Cipullo, S., Coulon, F., Pradhan, B., 2018. A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination. Sci. Total Environ. 644, 954–962.
Sayed, G.I., Hassanien, A.E., Azar, A.T., 2019. Feature selection via a novel chaotic crow search algorithm. Neural Comput. Appl. 31 (1), 171–188.
Sokolova, M., Japkowicz, N., Szpakowicz, S., 2006. In: December). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. Springer, Berlin, Heidelberg, pp. 1015–1021.
Song, Y., Qin, S., Qu, J., Liu, F., 2015. The forecasting research of early warning systems for atmospheric pollutants: a case in Yangtze River Delta region. Atmos. Environ. 118, 58–69.
Stafoggia, M., Bellander, T., Bucci, S., Davoli, M., de Hoogh, K., de' Donato, F., Gariazzo, C., Lyapustin, A., Michelozzi, P., Renzi, M., Scortichini, M., Shtein, A., Viegi, G., Kloog, I., Schwartz, J., 2019. Estimation of daily PM10 and PM2.5 contaminations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. Environ. Int. 124, 170–179.
Stafoggia, M., Samoli, E., Alessandrini, E., et al., 2013. Short-term associations between fine and coarse particulate matter and hospitalizations in Southern Europe: results from the MED-PARTICLES project. Environ. Health Perspect. 121, 1026–1033.
Statistical Institute of Catalonia (IDESCAT), (2019). Population on 1 January. Provinces. https://www.idescat.cat/pub/?id=aec&n=245&lang=en en.
Stull, R.B., 2012. An introduction to boundary layer meteorology Vol, 13)..
Suleiman, A., Tight, M.R., Quinn, A.D., 2019. Applying machine learning methods in managing urban contaminations of traffic-related particulate matter (PM10 and PM2. 5). Atmos. Pollution Res. 10 (1), 134–144.
Sutton, C.D., 2004. Classification and regression trees, bagging, and boosting. Handb. Stat. 24, 303–329. https://doi.org/10.1016/S0169-7161(04)24011-1.
Taspinar, F., 2015. Improving artificial neural network model predictions of daily average PM10 contaminations by applying principle component analysis and implementing seasonal models. J. Air Waste Manag. Assoc. 65, 800–809.
Timofeev, R., 2004. Classification and regression trees (CART) theory and applications. Humboldt University, Berlin.
Van Damme, M. et al., 2018. Nature 564, 99–103.
Voukantsis, D., Karatzas, K., Kukkonen, J., Räsänen, T., Karppinen, A., Kolehmainen, M., 2011. Intercomparison of air quality data using principal component analysis, and forecasting of PM 10 and PM 2 5 contaminations using artificial neural networks, in Thessaloniki and Helsinki. Sci Total Environ 409, 1266–1276.
Wang, D.Y., Wei, S., Luo, H.Y., Yue, C.Q., Grumder, O., 2017. A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine. Sci. Total Environ. 580, 719–733.
Liu, X.P., Ma, L., Li, X., Ai, B., Li, S.Y., He, Z.J., 2014. Simulating urban growth by integrating landscape expansion index (LEI) and cellular automata Int. J. Geogr. Inf. Sci. 28 (1), 148–163.
Yaseen, Z.M., Sulaiman, S.O., Deo, R.C., Chau, K.-W., 2019. An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical applications in water resource engineering area and future research direction. J. Hydrol. 569, 387–408. https://doi.org/10.1016/j.jhydrol.2018.11.069.
Zhang, H., Zhang, W.D., Palazoglu, A., Sun, W., 2012. Prediction of ozone levels using a hidden Markov model (HMM) with gamma distribution. Atmos. Environ. 62, 64–73.
Zhou, Y., Chang, F.J., Chang, L.C., Kao, I.F., Wang, Y.S., Kang, C.C., 2019. Multi-output support vector machine for regional multi-step-ahead PM2. 5 forecasting. Sci. Total Environ. 651, 230–240.
Zhu, S., Lian, X., Wei, L., Che, J., Shen, X., Yang, L., Li, J., 2018. PM2. 5 forecasting using SVR with PSOGSA algorithm based on CEEMD, GRNN and GCA considering meteorological factors. Atmos. Environ. 183, 20–32.