Ecotoxicology and Environmental Safety 182 (2019) 109386

Contents lists available at ScienceDirect

Ecotoxicology and Environmental Safety

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

Air quality warning system based on a localized $PM_{2.5}$ soft sensor using a novel approach of Bayesian regularized neural network via forward feature selection

Deepak Balram[a], Kuang-Yow Lian[a,*], Neethu Sebastian[b]

[a] Department of Electrical Engineering, National Taipei University of Technology, No. 1, Section 3, Zhongxiao East Road, Taipei, 106, Taiwan, Republic of China
[b] Institute of Organic and Polymeric Materials, National Taipei University of Technology, No. 1, Section 3, Zhongxiao East Road, Taipei, 106, Taiwan, Republic of China

ARTICLE INFO

Keywords:
Soft sensors
Pollutants
Environmental safety
$PM_{2.5}$
SVM classifier
Neural network

ABSTRACT

It is highly significant to develop efficient soft sensors to estimate the concentration of hazardous pollutants in a region to maintain environmental safety. In this paper, an air quality warning system based on a robust $PM_{2.5}$ soft sensor and support vector machine (SVM) classifier is reported. The soft sensor for the estimation of $PM_{2.5}$ concentration is proposed using a novel approach of Bayesian regularized neural network (BRNN) via forward feature selection (FFS). Zuoying district of Taiwan is selected as the region of study for implementation of the estimation system because of the high pollution in the region. Descriptive statistics of various pollutants in Zuoying district is computed as part of the study. Moreover, seasonal variation of particulate matter (PM) concentration is analyzed to evaluate the impact of various seasons on the increased levels of PM in the region. To investigate the linear dependence of concentration of different pollutants to the concentration of $PM_{2.5}$, Pearson correlation coefficient, Kendall's tau coefficient, and Spearman coefficient are computed. To achieve high performance for the $PM_{2.5}$ estimation, selection of appropriate forward features from the input variables is carried out using FFS technique and Bayesian regularization is incorporated to the neural network system to avoid the overfitting problem. The comparative evaluation of performance of BRNN/FFS estimation system with various other methods shows that our proposed estimation system has the lowest mean square error (MSE), root mean square error (RMSE), and mean absolute error (MAE). Moreover, the coefficient of determination (R-squared) is around 0.95 for the proposed estimation method, which denotes a good fit. Evaluation of the SVM classifier showed good performance indicating that the proposed air quality warning system is efficient.

1. Introduction

In this modern era, environmental pollution is a major concern to mankind. Industrialization is the prime reason for environmental pollution in both developed and developing countries. The presence of poor air quality can cause critical and chronic illness to human beings (Ning et al., 2019) (Bai et al., 2018) (Qiu et al., 2019) (Shou et al., 2019). The dense fog emerged over Greater London in 1952 for 4 days which resulted in more than 10,000 deaths is considered one of the primary reason for the initiative of monitoring the air quality of a region and carryout studies to have insight about the health issues caused by various pollutants in air (Logan, 1953) (Wu et al., 2017). Even though many countries have implemented various regulations to reduce air pollution, maintaining a good air quality is still a major concern for most economies. According to the latest data of World Health Organization (WHO), ambient air pollution has resulted in 4.2 million deaths every year worldwide and 91% of world's population is still living in regions where the air quality is over the WHO guideline limits (Jia et al., 2019). Hence it is significant to determine the air quality of a region and the data should be made available to public as a matter of safety concern. PM is extensively monitored to study the air quality of a region because of its dangerous environmental and health impacts. Investigations shows that PM adversely affects more people than any other pollutants and it has direct correlation to death rate and hospitalization cases because of cardiac and lung diseases (Wang et al., 2019) (Zhang et al., 2019) (Querol et al., 2001). Severe visibility reduction due to haze is another critical problem caused by the high concentration of PM in ambient air (Wang et al., 2017) and studies in Taiwan show that both $PM_{2.5}$ and PM_{10} have caused severe atmospheric visibility problems (Lee et al., 2005) (Tsai, 2005) (Tsai and Cheng, 1999). PM having an aerodynamic diameter less than 2.5 μm ($PM_{2.5}$) has gained significant attention as it can persist in the air for long time and

* Corresponding author.
E-mail address: kylian@ntut.edu.tw (K.-Y. Lian).

https://doi.org/10.1016/j.ecoenv.2019.109386
Received 5 April 2019; Received in revised form 22 June 2019; Accepted 24 June 2019
Available online 28 June 2019
0147-6513/

is considered very harmful to human beings (Vu et al., 2012) (Niu et al., 2015). It is composed of diverse hazardous chemicals (Yang et al., 2014) and can easily penetrate into lungs and cardiovascular system of human beings resulting in critical illness. Related studies in Taiwan show that exhausts from traffic and secondary aerosols including SO_4^{2-}, NO_3-, and NH_4+ are the major factors of $PM_{2.5}$ in ambient air (Chen et al., 2001). PM having an aerodynamic diameter less than 10 µm (PM_{10}) consists of heavy metals and the main contributing factors are vehicles, industrial emissions, and agricultural heating (Fang et al., 1999). Even though PM_{10} is not as dangerous as $PM_{2.5}$, related investigations show that long-term exposure to PM_{10} can cause cardiovascular and respiratory diseases (Mohamad et al., 2016) (Reche et al., 2012) (Zheng et al., 2019). It is noted that there is around 1% increase in the daily mortality rate for every 10 μgm^{-3} increase in PM_{10} concentration (Chen et al., 2003). Hence it is significant to estimate and evaluate PM concentrations in an area, especially $PM_{2.5}$.

Estimating the air quality of a region and issuing warning based on the air quality value is critical to avoid exposure to hazardous pollutants and thus avoid health issues. Soft sensors are an apt alternative to expensive traditional instrumentation for estimation of $PM_{2.5}$. The development of soft sensors is inexpensive as it can be developed using low cost hardware, namely microcontrollers. Another advantage of using soft sensor for $PM_{2.5}$ estimation instead of traditional instruments is that estimation based on soft sensors are comparatively faster (Lin et al., 2007). Moreover, many factors like the characteristic of the instruments, regional environmental issues etc. might influence the variance of the data measured based on traditional instruments. In this case soft sensors are a better alternative to traditional instruments. Different approaches can be used to model and estimate various air quality indexes and the choice of a method depends on the resources or data available and the complexity of the problem addressed. Deterministic air quality prediction methods are generally less stable and requires more time when compared with statistical methods (Cobourn, 2010) (Isukapalli, 1999) (Zhu et al., 2018). Particularly in the case of more complex estimations or modelling, statistical methods are the most desirable solution. Statistical methods including autoregressive integrated moving average model (ARIMA) (Arumugam and Saranya, 2018) (Zhou et al., 2014), hidden Markov models (HMM) (Sun et al., 2013), gray models (Wang and Hao, 2016), support vector regressions (SVRs) (Vapnik, 1999) (Lin et al., 2011), artificial neural networks (ANNs) (Zhou et al., 2014), multiple linear regression (MLR) (Sousa et al., 2007), and hybrid models (Sun and Sun, 2017) (Gan et al., 2018) are adopted by researchers to determine air pollution indexes. Among these methods, multiple linear regression is the widely used statistical prediction method because of its accuracy and less complexity. Prediction of NO_x and PM_{10} concentrations using the predictors NO, NO_2, CO, O_3 and $PM_{2.5}$ is carried out based on multiple linear regression by Vlachogianni (Vlachogianni et al., 2011). Non-linear regression approaches were also studied to estimate $PM_{2.5}$ value by researchers (Cobourn, 2010) (Baker and Foley, 2011). $PM_{2.5}$ concentrations is related to various factors and hence statistical estimations of $PM_{2.5}$ can be made based on the measurement of these related factors. Related investigations show that estimation of $PM_{2.5}$ from meteorological measures was carried out by researchers using non-linear exposure-lag-response model (Chen et al., 2018). Neural networks can be used to achieve non-linear statistical modelling and can make accurate estimation with comparatively less statistical training. The significant advantage of using neural networks to estimate $PM_{2.5}$ concentrations is that it has the capability to identify all possible interactions between input variables and it recognizes hidden nonlinear relationships between responses and input variables. Non-linear functions can be easily modelled and valid generalization can be made to accurately estimate the values if new test data is provided with the help of neural network (Huang et al., 2017; Ruan et al., 2017). In this work, we have used a novel approach of prediction technique based on neural network via FFS to estimate the $PM_{2.5}$ value. FFS helps the neural network to achieve high performance by selecting the best possible subset of input features. Apart from estimation, a warning system is critical in analyzing the air quality and take precautionary measures to protect ourselves from the adverse effects of air pollution. We have used SVM classifier to model the air quality warning system based on the concentration of estimated $PM_{2.5}$ and other pollutants.

This work mainly focuses on the significance of air quality estimation in Zuoying district of Taiwan and the novel approach is used to achieve precise determination of $PM_{2.5}$ concentration and modelling of air quality warning system. Section 2 of the paper details the methods and analysis used in this work. The importance of selecting Zuoying district of Taiwan for this study, the air quality data and dataset preparation, the importance of all the input variables used for the estimation of $PM_{2.5,}$ descriptive statistics of various pollutants, and seasonal variation of PM in Zuoying district are analyzed in this section. Moreover, $PM_{2.5}$ estimation based on BRNN/FFS technique and the modelling of air quality warning system based on SVM classifier are also explained in detail in this section. The results obtained from this work and discussion are provided in section 3 and section 4 concludes the paper.

2. Methods and analysis

2.1. Region of study

Taiwan is one of the most densely populated country with 652 persons per square kilometer as per the latest data of February 2019 by the statistical bureau of Taiwan. Many regions in Taiwan is currently facing alarming air pollution consequences because of the high level of particulate matter in ambient air. The high population density, domestic industries and automobile exhaust emissions are the major contributing factors of air pollution in Taiwan (Chang et al., 2005) (Yang et al., 2004) (Chen et al., 1999). According to Environmental Protection Administration (EPA) of Taiwan, air quality index (AQI) value beyond 50 is considered unhealthy and during winter season, the air quality of southern and central Taiwan often registers in the red zone, indicating very poor and unsafe air quality. According to EPA's Department of Environmental Monitoring and Information Management, China is a major source of $PM_{2.5}$ in Taiwan's atmosphere because of the geographic position. China has made immense progress in industrialization and it is noted that the strong winds carry the pollutants from industrial areas in China and this contribute an increase in the $PM_{2.5}$ concentration of Taiwan (Wu et al., 2004) (Tsai and Chen, 2006). Taiwan's topography also has an adverse influence to air quality of the region as it is surrounded by high mountains and hence the local pollutants tend to accumulate in the region if there is no wind. Hazardous pollutants like sulfur dioxide (SO_2), nitrogen dioxide (NO_2), ground-level ozone (O_3), carbon monoxide (CO), and volatile organic compounds (VOCs) primarily derived from the burning of fossil fuels contribute to poor air quality of Taiwan. Scientific studies show that exposure to $PM_{2.5}$ is associated with adverse health problems in Taiwan including chronic obstructive pulmonary disease (COPD), lung function decline, ischemic heart disease, cancer, asthma, and pneumonia (Guo et al., 2018) (Kuo et al., 2002) (Tsai and Yang, 2014) (Chiu et al., 2013) (Hung et al., 2012).

Kaohsiung is the third most populous city in Taiwan and is considered as an industrial center of southern Taiwan as many factories and power plants are located in the city. Our study is focused on estimation of air quality in Zuoying district in Kaohsiung city. Fig. 1 depicts the map showing the area of our study and is created using QGIS software. Zuoying belongs to the inner Kaohsiung region and has a population of 196,362 as per 2016 census. According to the report of the environmental group Air clean Taiwan, Zuoying district has the worst air pollution in Taiwan in terms of $PM_{2.5}$ level in the year 2016. Therefore, it is significant to conduct a study to analyze the air quality of Zuoying district. The population density of Zuoying is 10,127 persons

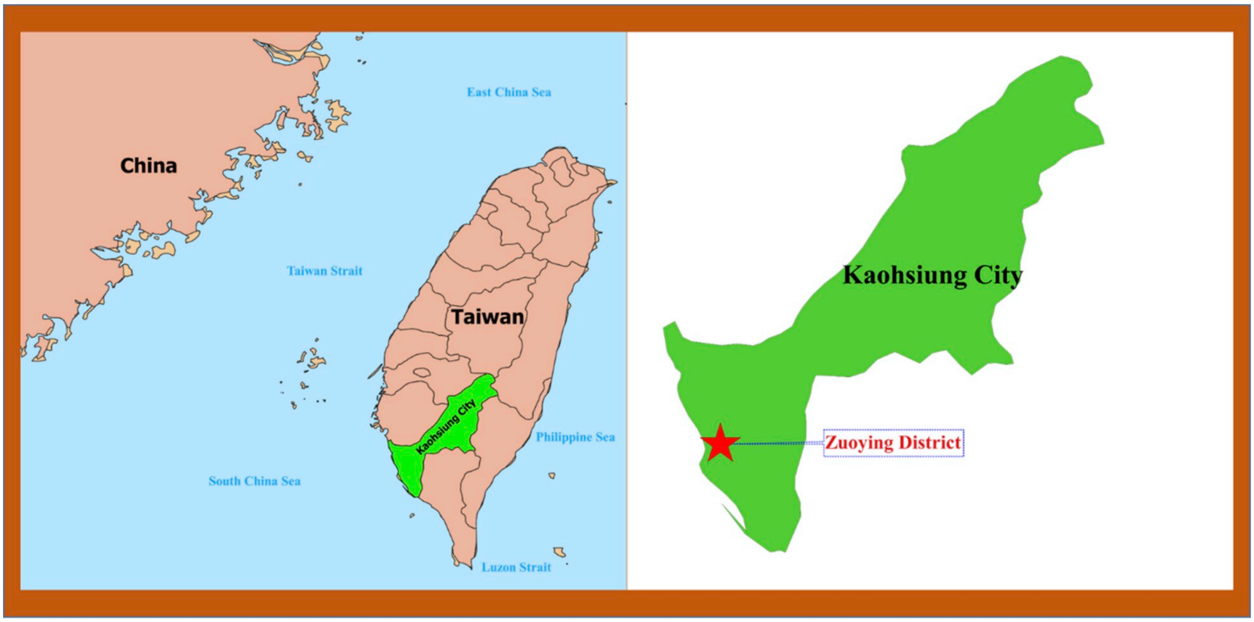

Fig. 1. Map showing area of study.

per square kilometer, which is almost 15 times the population density of Taiwan. The Fushan village, which is the most populous village in Taiwan is also located in Zuoying district. High population density is one of the major contributing factor to worst air quality in Zuoying. Another significant reason why air pollutants get accumulated in Zuoying is because of the geographical positioning of the region. The wind speed in this region is comparatively less during winter as it is located against the wind direction in the western side of central mountain range. As a result, the pollutants tend to accumulate in the ambient air of the region as it is not carried away by the wind causing deterioration in the air quality of the region. The biggest naval base of Taiwan and Zuoying naval airfield are located in Zuoying district. Studies shows that emissions from ship can adversely affect the air quality in coastal areas as the concentration of pollutants including CO_2, CO, NO_x, SO_2 etc. increases (Fang et al., 2006) (Eckhardt et al., 2013) (Tian et al., 2018). Further investigations have proved that ship emissions can cause an increase in $PM_{2.5}$ concentrations in a region and can lead to premature deaths due to lung cancer and cardiovascular diseases (Matthias et al., 2010). Aviation fuel combustion also deteriorates the air quality of a region as pollutants including PM, NO, CO and hydrocarbons are released to atmosphere (Brzozowski and Kotlarz, 2005). So the naval airfield and naval base in Zuoying district can also be a reason for the worst air pollution in the region.

2.2. Air quality data analysis

The datasets for this work is collected and prepared using the data from environmental protection administration of Taiwan. We have taken the hourly data of pollutants including PM_{10}, $PM_{2.5}$, CO, NO_2, SO_2, and O_3 pertaining to Zuoying district for preparing the dataset used for the estimation of $PM_{2.5}$ and modelling of air quality warning system. Every input variables and target variable has 1587 data points and we have used 70% of the samples for training and the rest of the data for testing and validation purpose. The division of data was carried out based on dividerand function using random indices. Data preprocessing is an important step in data-driven soft sensors as incorrect input data can result in inaccurate model. Moreover, the characteristic of the instruments, regional environmental factors and other factors might influence the variance of the data collected by EPA, Taiwan and hence data preprocessing is a necessity in our work. From the hourly data of pollutants collected from EPA, those observations that deviates remarkably from normal values and missing data are considered as outliers and are removed in the data preprocessing stage to achieve accurate dataset for this study. Removal of these outliers is considered as an important step in the development of soft sensors. Studies shows that seasons have impact on the concentration of pollutants in an area and hence we have taken pollutant data of four months of the year 2018 (January, April, July, and October) representing four seasons of Taiwan mainly winter, spring, summer, and autumn respectively for creating the estimation dataset.

To analyze the effect of seasons in PM concentration of Zuoying district, we have computed the descriptive statistics of PM_{10} and $PM_{2.5}$ during different seasons in Taiwan and given in Table 1. The mean, mode and standard deviation are computed to have insight about average concentration, most frequent value of each pollutant and variations in the concentration of pollutants. When we analyze Table 1, we can clearly conclude that seasons have a great impact on the concentration of PM in Zuoying district. The 24-h mean air quality guideline values of WHO for pollutants PM_{10}, and $PM_{2.5}$ are 50 μg/m^3, and 25 μg/m^3 respectively. The mean PM_{10} concentration during winter, autumn, and spring seasons are 83.51 μg/m^3, 67.81 μg/m^3, and 61.12 μg/m^3 respectively, which are all above the WHO guideline limits. Moreover, the mean $PM_{2.5}$ concentrations 39.74 μg/m^3, 29.25 μg/m^3, and 30.85 μg/m^3 during winter, autumn, and spring seasons respectively are also above the WHO guideline limits. The most frequent PM_{10} and $PM_{2.5}$ concentrations in the Zuoying area during winter season, 67 μg/m^3 and 33 μg/m^3 respectively are also very high. So comparatively, summer season is better in terms of air quality based on PM concentration and the worst being winter season.

Correlation is a bivariate analysis and the correlation coefficient

Table 1
Seasonal variation in PM concentration.

Seasons	PM_{10} (μg/m^3)			$PM_{2.5}$ (μg/m^3)		
	Mean	Mode	Standard Deviation	Mean	Mode	Standard Deviation
Spring	61.12	53	26.95	30.85	16	17.38
Summer	35.58	32	14.77	12.88	8	7.89
Autumn	67.81	62	35.58	29.25	26	17.10
Winter	83.51	67	33.12	39.74	33	18.42

value varies between −1 and +1, where a value close to +1 or −1 indicates a strong association and a value close to 0 indicates a weaker association. The Pearson correlation coefficient, Kendall's tau coefficient, and Spearman coefficient are computed to investigate the linear dependence of concentration of different pollutants including SO_2, CO, O_3, NO_2, and PM_{10} to the concentration of $PM_{2.5}$. Among all the pollutants considered, PM_{10} returned a value closest to 1 for all the aforementioned correlation coefficients indicating that PM_{10} has the strongest association to $PM_{2.5}$. Moreover, O_3 returned a value closest to 0 indicating that it has comparatively the weakest association to $PM_{2.5}$. Also, we have got p-value of 0 for all the pollutant variables indicating that the input variables considered are statistically significant for the estimation of $PM_{2.5}$. For column A_x in matrix A and column B_y in matrix B, equations (1)–(3) given below denotes Pearson coefficient, Kendall's tau coefficient, and Spearman coefficient respectively (Bolboaca and Jäntschi, 2006). In the below equations, n denotes the length of each column and d denotes the distance between the rank of 2 columns.

$$\text{rho}(x, y) = \frac{\sum_{i=1}^{n}(A_{x,i} - \bar{A}_x)(B_{y,i} - \bar{B}_y)}{\left\{\sum_{i=1}^{n}(A_{x,i} - \bar{A}_x)^2 \sum_{j=1}^{n}(B_{y,j} - \bar{B}_y)^2\right\}^{1/2}} \tag{1}$$

$$\tau = \frac{2K}{n(n-1)}, \text{ where } K = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \xi^*(A_{x,i}, A_{x,j}, B_{y,i}, B_{y,j}) \text{ and } \xi^*(A_{x,i}, A_{x,j}, B_{y,i}, B_{y,j}) = \begin{cases} 1 \text{ if } (A_{x,i} - A_{x,j})(B_{y,i} - B_{y,j}) > 0 \\ 0 \text{ if } (A_{x,i} - A_{x,j})(B_{y,i} - B_{y,j}) = 0 \\ -1 \text{ if } (A_{x,i} - A_{x,j})(B_{y,i} - B_{y,j}) < 0 \end{cases} \tag{2}$$

$$\text{rho}(x, y) = 1 - \frac{6\sum d^2}{n(n^2 - 1)} \tag{3}$$

The response plot of concentration of various pollutants including SO_2, CO, O_3, NO_2, and PM_{10} towards the concentration of $PM_{2.5}$ is given in Fig. 2. When we visualize Fig. 2, it is clear that the concentration of PM_{10} is comparatively more linear towards the concentration of $PM_{2.5}$ indicating that PM_{10} concentration will play the major role than the concentration of other pollutants in the estimation of $PM_{2.5}$ concentration.

2.3. Estimation of $PM_{2.5}$ based on BRNN/FFS

The estimation of $PM_{2.5}$ concentration was carried out based on a Bayesian regularized neural network via FFS. Neural network has the capability to recognize complex nonlinear relationships between responses and input variables and also determine all feasible associations between predictor variables (Tu, 1996). Modelling of non-linear functions is less complicated and valid generalization can be made to accurately estimate the values if new test data is provided with the help of neural network. Good generalization is a key necessity for a good neural network system and one major issue faced by neural network is the overfitting problem. In this context, regularization is an important process to reduce the overfitting problems and hence achieve efficient neural network system (Srivastava et al., 2014). So in our proposed $PM_{2.5}$ estimation system based on neural network, we have incorporated a Bayesian regularization to avoid the overfitting problem and whereby achieve accurate estimation results. Bayesian regularization helps to reduce a linear combination of network weights and squared errors of neural network. The objective function parameters should be assigned exact values and is key in regularization (Burden and Winkler, 2008).

In our work, as we are incorporating Bayesian regularization, network weights are considered as random variables and the density functions of the weights are updated based on Bayes' theorem as mentioned below in equation (4) (MacKay, 1992):

$$P(W|S, \alpha, \beta, N) = \frac{P(S|W, \beta, N)P(W|\alpha, N)}{P(S|\alpha, \beta, N)} \tag{4}$$

where W denotes network weights, S represents the dataset, α and β are the hyper parameters, N denotes the neural network model used, $P(S|W, \beta, N)$ represents the likelihood function, $P(S|\alpha, \beta, N)$ is the normalization factor, and $P(W|\alpha, N)$ denotes the prior density. Fig. 3

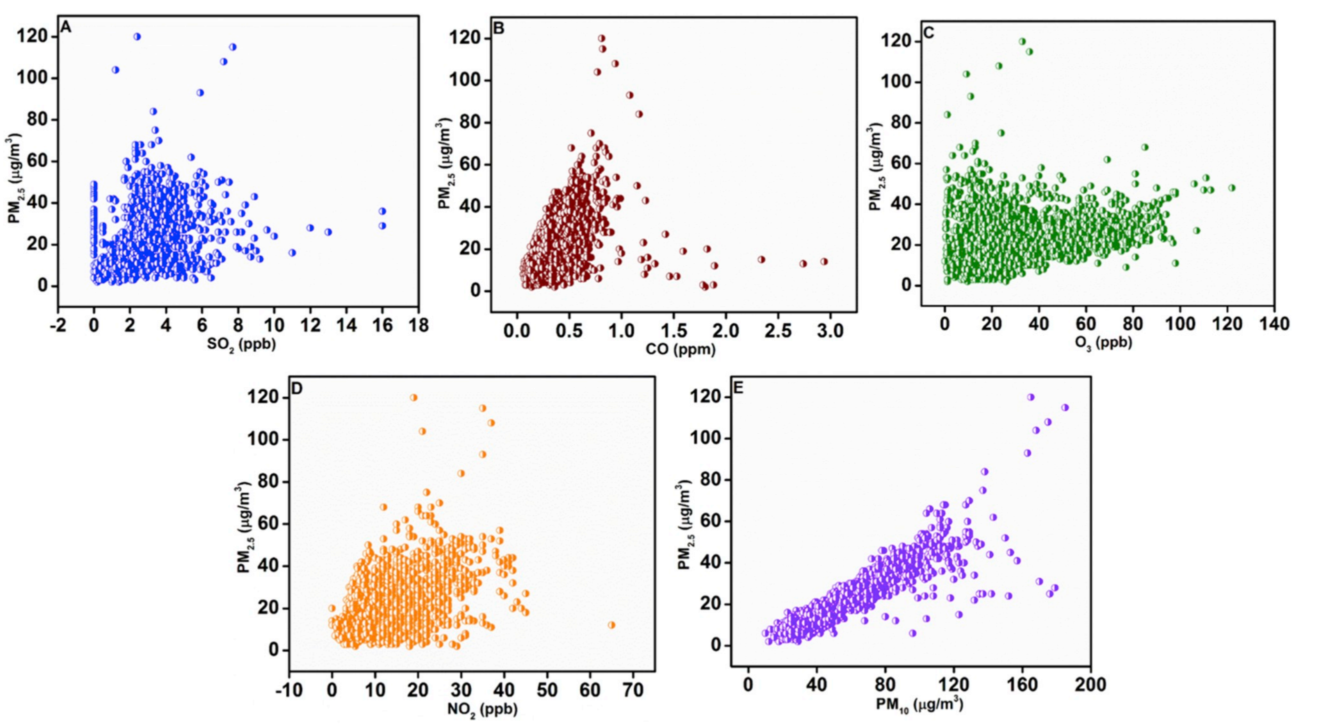

Fig. 2. Response plot of concentration of various pollutants: SO_2 (A); CO (B); O_3 (C); NO_2 (D); PM_{10} (E) vs concentration of $PM_{2.5}$.

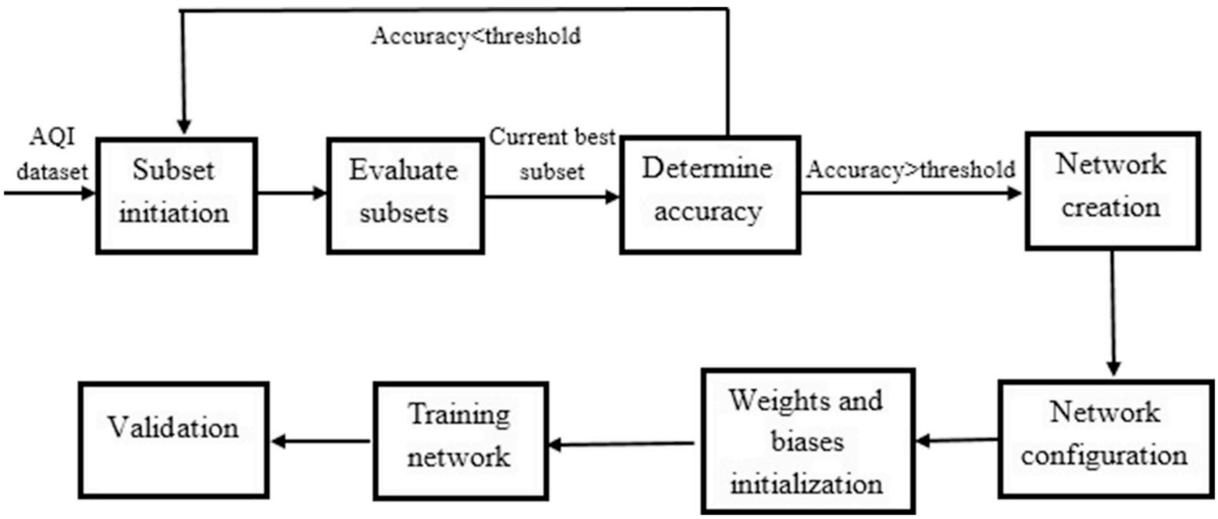

Fig. 3. Block diagram of $PM_{2.5}$ estimation based on neural network via FFS.

shows the block diagram of $PM_{2.5}$ estimation based on BRNN/FFS. A three-layer neural network architecture with 10 neurons in the hidden layer and 1 neuron in the output layer is used for the estimation of $PM_{2.5}$ in this work.

In this approach, selection of appropriate forward features from the input variables is carried out using FFS technique to achieve high performance for the estimation. FFS will aid to reduce the overfitting problem and low efficiency of learning due to unnecessary features, often regarded as dimensionality curse for BRNN (Jain and Zongker, 1997). FFS will return the best subset of input variables that can be further fed to the neural network system to achieve high performance. In the first stage of this proposed method of BRNN via FFS, we start by selecting one input variable from the dataset as the subset and evaluate the performance. This process is repeated for each input variable and the variable that offers the best performance is selected and appended to the relevant input variable list. After the selection of the best input variable from the dataset, the aforementioned subset initiation process is reiterated with two input variables, one from the input variables not selected in the first subset evaluation process and the other one is the input variable selected in the first stage of subset evaluation. The set of input variables form a subset and the performance is evaluated for each set of input variables created. The input variable that offers the best performance in this second stage iteration is appended to the relevant input variable list. This process is repeated by adding the input variable which best improves our model in each iteration until we achieve the intended performance, set as the threshold value. In this way we could tune the input variables to meet optimal performance for the neural network. Apart from the performance factor, another important advantage of this proposed methodology is the reduction in complexity of the neural network system and hence comparatively easier interpretation of the system (Liu and Motoda, 2012).

The performance of the proposed estimation system was evaluated to determine the efficiency of the $PM_{2.5}$ concentration determination. Performance evaluation was carried out by calculating MSE, MAE, RMSE, and R-squared value. Equations (5)–(8) given below denote the calculation of MSE, RMSE, MAE, and R-squared value respectively:

$$\text{MSE}=\frac{1}{n}\sum_{i=1}^{n}(p_i-\tilde{p}_i)^2 \tag{5}$$

$$\text{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i-\tilde{p}_i)^2} \tag{6}$$

$$\text{MAE}=\frac{1}{n}\sum_{i=1}^{n}|p_i-\tilde{p}_i| \tag{7}$$

$$\text{R}-\text{ squared}=1-\frac{\sum_{i=1}^{n}(p_i-\tilde{p}_i)^2}{\sum_{i=1}^{n}(p_i-\bar{p}_i)^2} \tag{8}$$

where n is the number of data points, p_i is the true value, $\tilde{p}_i$ is the estimated value, and $\bar{p}_i$ is the mean value.

2.4. Modelling air quality warning system

For modelling the air quality warning system, we have used a linear classifier, namely SVM. SVM is a binary classifier and it make use of geometric criteria instead of statistical information for classification. So it doesn't require a more complicated calculation of statistical distribution in order to carryout classification task. The fundamental technique of this classifier is the implementation of optimal hyperplane algorithm (Barbosa et al., 2016). Optimal hyperplane is regarded as the plane that provides maximum margin of dissociation between the target classes. Consider the dataset for training is $S = \{(P_1, q_1), \ldots, (P_k, q_k)\}$, we have to solve the constrained optimization problem represented below in equation (9) in order to determine the optimal hyperplane depicted in equation (10) (Melgani and Bruzzone, 2004) (Maione et al., 2018),

$$\min_{w}\frac{1}{2}||P||^2 \text{subject to qi((W·Pi + a))} \geq 1 \text{ where } i = 1, \ldots, k \tag{9}$$

$$\text{q(P)} = \text{W} \cdot \text{P} \tag{10}$$

Based on the Lagrangian expression, the linear optimization problem can be written using Lagrange multipliers α_i as the following equation:

$$\max_{\alpha}\sum_{i=1}^{k}\alpha_i-\frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k}\alpha_i\alpha_j q_i q_j(P_i\cdot P_j)\text{subject to }\alpha_i \geq 0, \sum_{i=1}^{k}\alpha_i q_i=0, \text{ i= 1, …,k} \tag{11}$$

Ultimately, the SVM implements the decision function described below:

$$\text{f (P)} = \text{sgn}\left(\sum_{i=1}^{k}\alpha_i q_i(P_i\cdot P)+\text{a}\right) \tag{12}$$

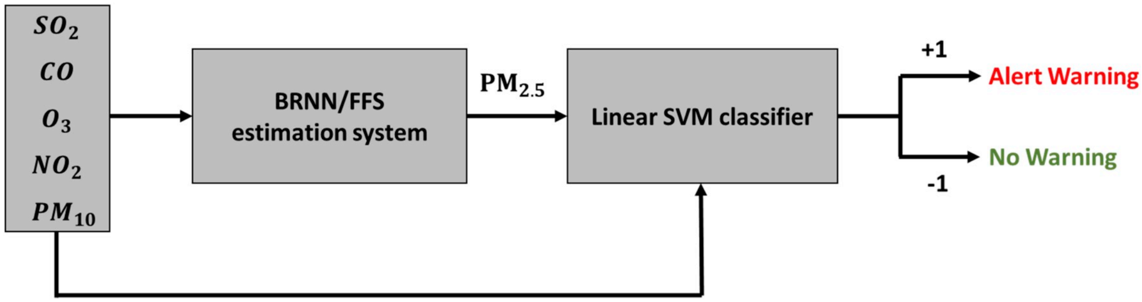

Fig. 4. Block diagram of air quality warning system.

Table 2
Comparative evaluation of the performance of the proposed estimation system.

Algorithm	MSE	RMSE	MAE	R-squared
Linear SVM	33.686	5.8039	3.6078	0.82
Linear regression	33.349	5.7749	3.635	0.82
Fine tree	23.203	4.817	3.3663	0.88
Bagged trees ensemble	20.959	4.5781	2.8934	0.89
Quadratic SVM	17.73	4.2107	2.947	0.91
Boosted trees ensemble	17.383	4.1693	2.971	0.91
Gaussian process regression	8.6973	2.9491	2.3971	0.95
BRNN	8.347	2.8891	2.3652	0.95
BRNN/FFS	7.4972	2.7381	2.3292	0.95

Fig. 4 shows the block diagram of the proposed air quality warning system. As shown in Fig. 4, the $PM_{2.5}$ concentration estimated based on BRNN/FFS estimation system is fed as one of the input to SVM classifier. Then the SVM classifier is trained over the feature vectors including concentration of CO, NO_2, SO_2, O_3, PM_{10}, and $PM_{2.5}$ for the development of air quality warning system. The class allocation for the SVM classifier is carried out based on the criterion given below, which is developed according to the national ambient air quality standard set by EPA, Taiwan for various pollutants.

Class $+1$: (SO_2 concentration ≥ 36 ppb) OR (CO concentration ≥ 4.5 ppm) OR (NO_2 concentration ≥ 54 ppb) OR (O_3 concentration ≥ 0.055 ppm) OR (PM_{10} concentration $\geq 55\,\mu g/m^3$) OR ($PM_{2.5}$ concentration $\geq 15.5\,\mu g/m^3$).

Class -1: (SO_2 concentration < 36 ppb) AND (CO concentration < 4.5 ppm) AND (NO_2 concentration < 54 ppb) AND (O_3 concentration < 0.055 ppm) AND (PM_{10} concentration $< 55\,\mu g/m^3$) AND ($PM_{2.5}$ concentration $< 15.5\,\mu g/m^3$).

If the SVM classifier returns class $+1$ as the output, then an alert warning is generated by the air quality warning system indicating that the quality of air is bad and if the classifier returns class -1 as the output, then no warning is generated by the air quality system indicating that the quality of air is good.

A total of 1587 observations of each predictor were used for the binary classification based on SVM for modelling the air quality warning system. Six predictors were used in order to classify the air quality of Zuoying district to two response classes and issue alert warning accordingly. The performance of the classifier was evaluated by computing accuracy, precision, recall, specificity, and F1-score based on the equations given below (Sprague et al., 2014):

$$\text{Accuracy} = \frac{Correct\ classifications}{Number\ of\ observations} \times 100 \tag{13}$$

$$\text{Recall} = \frac{True\ positive}{True\ positive + False\ negative} \tag{14}$$

$$\text{Specificity} = \frac{True\ negative}{True\ negative + False\ positive} \tag{15}$$

$$\text{Precision} = \frac{True\ positive}{True\ positive + False\ positive} \tag{16}$$

$$\text{F1} - \text{score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{17}$$

3. Results and discussion

Selecting the number of hidden neurons for a neural network system is an important task as wrong selection can lead to either underfitting or overfitting issues, which can indeed adversely affect the overall efficiency of the system. Hence we have conducted a study to fix the number of hidden neurons to be used in the proposed BRNN/FFS system. We achieved the lowest RMSE when the number of hidden neurons was fixed at 10 and also the processing time was around 4.3639 s only. Hence we have used 10 hidden neurons in the BRNN/FFS estimation system. Table 2 compares the performance of the proposed BRNN/FFS estimation system with other algorithms. Analyzing the table, we can clearly understand that MSE, RMSE, and MAE are comparatively the lowest for our proposed system, which indicates the high efficiency of $PM_{2.5}$ estimation based on our method. Moreover, the coefficient of determination is around 0.95 for the proposed estimation method indicating a good fit.

Fig. 5A shows the performance graph of the BRNN/FFS estimation system. The estimation system achieves the lowest MSE at an epoch of 175 and it is at this point, the best performance of the system is realized. Fig. 5B shows the error histogram of the of the BRNN/FFS estimation system. The entire error range is divided into 20 bins and the vertical black line at the center of the histogram indicates zero error with around 185 instances in training and 35 instances in testing. We can find that the training and testing instances having error value greater than 7 is almost negligible indicating that the deviation of estimated $PM_{2.5}$ concentration from its true value is very small in almost all instances of training and testing.

Fig. 5C depicts the graph representing the $PM_{2.5}$ concentration estimated based on the BRNN/FFS system and the true value. By visualizing Fig. 5C, we can find that almost all the $PM_{2.5}$ concentration values estimated are close to true value and there is only negligible difference in the estimated value and true value. This indicates that the proposed estimation system based on BRNN/FFS is efficient. The aforementioned true value here indicates the results from traditional instrumental analysis approach. So the results from Fig. 5C indicates that the estimated $PM_{2.5}$ concentration based on low-cost and flexible BRNN/FFS is almost similar to that of expensive traditional instrumental analysis method. The traditional instrumental analysis method involves the use of expensive instruments and there will be difficulty to acquire results during unfavorable conditions. The regional environmental conditions can adversely impact the accuracy of the pollutant values estimated based on instrumental analysis method. In these circumstances, the proposed estimation system is a good alternative to get faster and accurate results as our system has very low RMSE of 2.73 $\mu g/m^3$. The regression curve of the estimation system is represented in

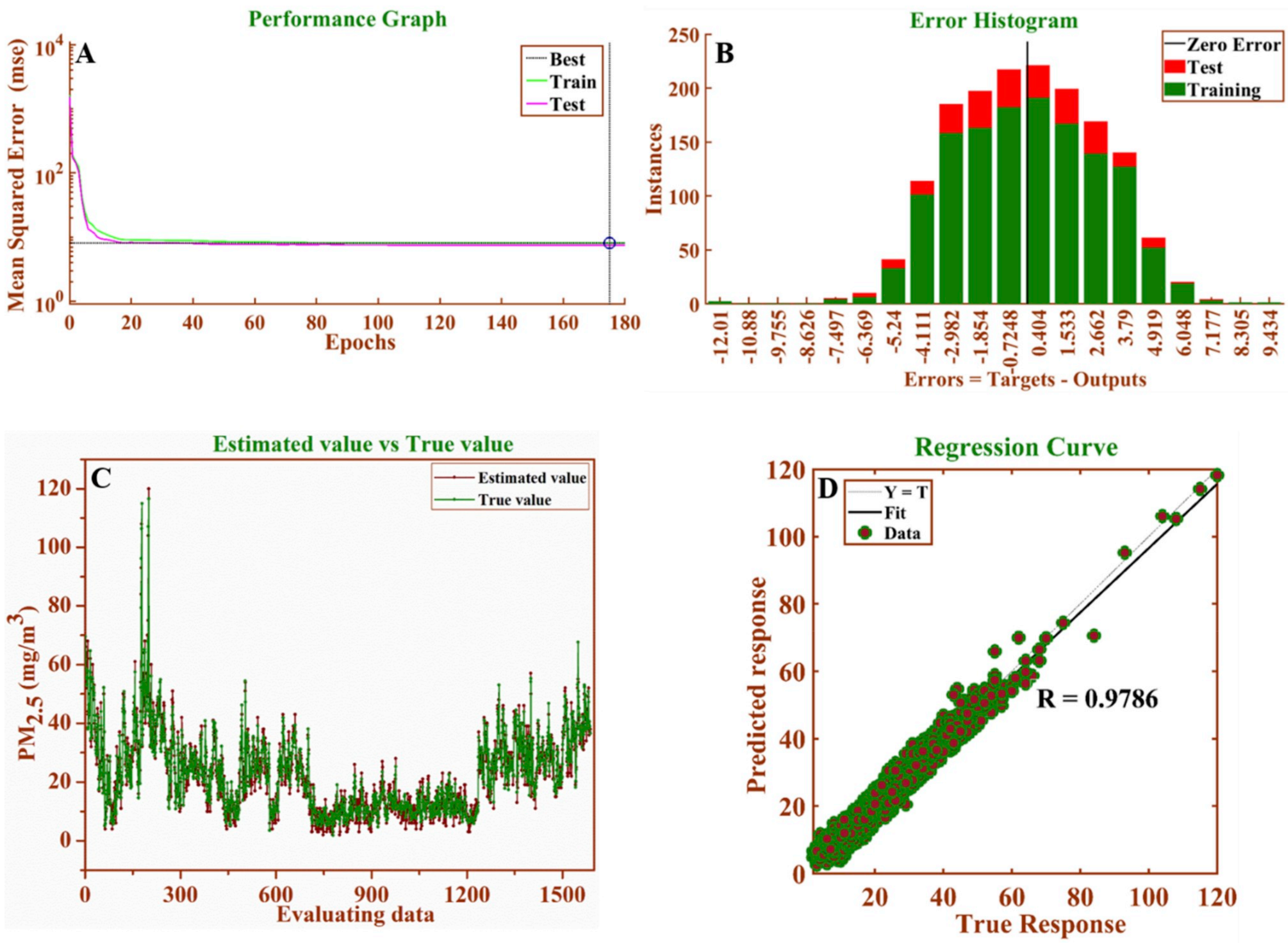

Fig. 5. Performance graph of the BRNN/FFS estimation system (A), error histogram (B), estimated value vs true value (C), and regression curve (D).

Fig. 5D and it shows that the R value (coefficient of correlation) of predicted response and true response is 0.9786 which is close to 1 indicating that the model prediction is very precise in this work.

Among the total of 1587 data points used for SVM classifier, 1555 data points are classified correctly and there were 32 misclassifications. Out of the 1555 correct classifications, there were 984 true positives and 571 true negatives. In the 32 misclassifications, 21 were false positives and 11 were false negatives. We have obtained an accuracy of 98% and F1-score of 0.983 for the air quality warning system based on BRNN/FFS and SVM classifier. All the performance metrics we evaluated returned perfect values indicating that the proposed air quality warning system is efficient.

Only few studies were attempted to estimate $PM_{2.5}$ in Taiwan based on different machine learning approaches using various factors. Jung et al. estimated ground-level $PM_{2.5}$ concentrations in Taiwan using long-term satellite based aerosol optical depth, localized land use data, and meteorological variables. The highest fitting achieved for the estimation model was 0.77 with a RMSE of 11.4 μg/m^3 (Jung et al., 2018). Also, Wu et al. assessed $PM_{2.5}$ concentrations in Taiwan based on a hybrid kriging/land-use regression model. The model was developed using data collected from 71 EPA monitoring stations from 2006 to 2011. The highest R-squared value achieved for the regression model was 0.88 with an RMSE of 5.02 μg/m^3 (Wu et al., 2018). When we consider similar studies carried out for $PM_{2.5}$ estimation in different places other than Taiwan, we can find that Wang et al. estimated $PM_{2.5}$ concentration using deep neural network (DNN) in Beijing–Tianjin–Hebei, China. In this study, meteorological parameters and gaseous pollutant concentrations were considered as predictor variables in order to estimate $PM_{2.5}$ concentration. The R-squared value achieved for the estimation of $PM_{2.5}$ was 0.87 and the RMSE was 27.11 μg/m^3 (Wang and Sun, 2019). Furthermore, Liu et al. conducted study based on random forest approach to estimate $PM_{2.5}$ concentration from reflectance at the top of the atmosphere. The model resulted in a R-squared value of 0.86 and an RMSE of 17.3 μg/m^3 for hourly $PM_{2.5}$ concentration estimation (Liu et al., 2019). Gaussian process method was employed by Yu et al. and geographically-weighted gradient boosting machine (GW-GBM) method was used by Zhan et al. for the estimation of $PM_{2.5}$ resulting in RMSE value of 21.87 μg/m^3 ($R^2 = 0.81$) and 23 μg/m^3 ($R^2 = 0.76$) respectively (Yu et al., 2017) (Zhan et al., 2017). When we make a comparative evaluation of our result with all the aforementioned studies, we can find that we are able to achieve an accurate estimation of $PM_{2.5}$ concentration in this work with a relatively low RMSE of 2.73 μg/m^3 and high R-squared value of 0.95. This indicates the high performance of the $PM_{2.5}$ estimation based on BRNN/FFS approach.

4. Conclusion

Determination of $PM_{2.5}$ concentration is significant as it is considered very harmful to mankind and its estimation will aid to get a clear idea about the air quality of a region. Zuoying district of Taiwan has the worst air pollution due to high concentration of $PM_{2.5}$ in the region and hence study of air quality of this region is important. Descriptive statistics of various pollutants in Zuoying district and its seasonal variation is analyzed based on the dataset prepared from EPA, Taiwan. The analysis of correlation coefficients of concentration among various pollutants shows that PM_{10} is highly correlated to $PM_{2.5}$ concentration. The proposed novel approach of BRNN/FFS in the development of $PM_{2.5}$ soft sensor resulted in accurate and less complex estimation of $PM_{2.5}$ concentration. The study conducted to evaluate the performance of BRNN/FFS system shows that the proposed method has achieved the lowest MSE, MAE, and RMSE than other estimation methods and the calculated R-squared value shows good fit. The modelling of air quality warning system is carried out based on SVM approach to achieve high efficiency and quick response, which are important factors when we consider warning systems. We have achieved accurate results for BRNN/FFS estimation and SVM classifier indicating that the proposed air quality warning system is efficient.

Acknowledgments

We thank the Environmental Protection Administration (EPA), Taiwan for providing air pollutant monitoring data. This work was supported by the Ministry of Science and Technology, ROC, under Grants 105–2221-E–027-061.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoenv.2019.109386.

References

Arumugam, P., Saranya, R., 2018. Outlier detection and missing value in seasonal ARIMA model using rainfall data*. Mater. Today: Proceedings 5, 1791–1799.
Bai, X., et al., 2018. Ultrafine particle libraries for exploring mechanisms of PM2.5-induced toxicity in human cells. Ecotoxicol. Environ. Saf. 157, 380–387.
Baker, K.R., Foley, K.M., 2011. A nonlinear regression model estimating single source concentrations of primary and secondarily formed PM2.5. Atmos. Environ. 45, 3758–3767.
Barbosa, R.M., et al., 2016. Recognition of organic rice samples based on trace elements and support vector machines. J. Food Compos. Anal. 45, 95–100.
Bolboaca, S.-D., Jäntschi, L., 2006. Pearson versus Spearman, Kendall's tau correlation analysis on structure-activity relationships of biologic active compounds. Leonardo J. Sci. 5, 179–200.
Brzozowski, K., Kotlarz, W., 2005. Modelling of air pollution on a military airfield. Atmos. Environ. 39, 6130–6139.
Burden, F., Winkler, D., 2008. Bayesian Regularization of Neural Networks. Artificial Neural Networks. Springer, pp. 23–42.
Chang, C.-C., et al., 2005. Air pollution and hospital admissions for cardiovascular disease in Taipei, Taiwan. Environ. Res. 98, 114–119.
Chen, K.-S., et al., 2001. Determination of source contributions to ambient PM2.5 in Kaohsiung, taiwan, using a receptor model. J. Air Waste Manag. Assoc. 51, 489–498.
Chen, M.-L., et al., 1999. The PM2.5 and PM10 particles in urban areas of Taiwan. Sci. Total Environ. 226, 227–235.
Chen, S.-J., et al., 2003. Characterization of atmospheric PM10 and related chemical species in southern Taiwan during the episode days. Chemosphere 53, 29–41.
Chen, Z.-Y., et al., 2018. Estimating PM2.5 concentrations based on non-linear exposure-lag-response associations with aerosol optical depth and meteorological measures. Atmos. Environ. 173, 30–37.
Chiu, H.-F., et al., 2013. Short-term effects of fine particulate air pollution on ischemic heart disease hospitalizations in taipei: a case-crossover study. Aerosol and Air Quality Research 13, 1563–1569.
Cobourn, W.G., 2010. An enhanced PM2.5 air quality forecast model based on nonlinear regression and back-trajectory concentrations. Atmos. Environ. 44, 3015–3023.
Eckhardt, S., et al., 2013. The influence of cruise ship emissions on air pollution in Svalbard – a harbinger of a more polluted Arctic? Atmos. Chem. Phys. 13, 8401–8409.
Fang, G.-C., et al., 1999. Characterization of chemical species in PM2.5 and PM10 aerosols in suburban and rural sites of central Taiwan. Sci. Total Environ. 234, 203–212.
Fang, G.-C., et al., 2006. Concentrations of ambient air particulates (TSP, PM2.5 and PM2.5–10) and ionic species at offshore areas near Taiwan Strait. J. Hazard Mater. 132, 269–276.
Gan, K., et al., 2018. A secondary-decomposition-ensemble learning paradigm for forecasting PM2.5 concentration. Atmos. Pollut. Res. 9, 989–999.
Guo, C., et al., 2018. Effect of long-term exposure to fine particulate matter on lung function decline and risk of chronic obstructive pulmonary disease in Taiwan: a longitudinal, cohort study. The Lancet Planetary Health 2, e114–e125.
Huang, M., et al., 2017. A new efficient hybrid intelligent model for biodegradation process of DMP with fuzzy wavelet neural networks. Sci. Rep. 7, 41239.
Hung, L.-J., et al., 2012. Traffic air pollution and risk of death from breast cancer in taiwan: fine particulate matter (PM2.5) as a proxy marker. Aerosol and Air Quality Research 12, 275–282.
Isukapalli, S.S., 1999. Uncertainty Analysis of Transport-Transformation Models. Rutgers University.
Jain, A., Zongker, D., 1997. Feature selection: evaluation, application, and small sample performance. IEEE Trans. Pattern Anal. Mach. Intell. 19, 153–158.
Jia, J., et al., 2019. Cr(VI)/Pb2+ are responsible for PM2.5-induced cytotoxicity in A549 cells while pulmonary surfactant alleviates such toxicity. Ecotoxicol. Environ. Saf. 172, 152–158.
Jung, C.-R., et al., 2018. Incorporating long-term satellite-based aerosol optical depth, localized land use data, and meteorological variables to estimate ground-level PM2. 5 concentrations in Taiwan from 2005 to 2015. Environ. Pollut. 237, 1000–1010.
Kuo, H.W., et al., 2002. Respiratory effects of air pollutants among asthmatics in central taiwan. Arch. Environ. Health 57, 194–200.
Lee, C.-G., et al., 2005. Effects of aerosol species on atmospheric visibility in Kaohsiung city, Taiwan. J. Air Waste Manag. Assoc. 55, 1031–1041.
Lin, B., et al., 2007. A systematic approach for soft sensor development. Comput. Chem. Eng. 31, 419–425.
Lin, K.-P., et al., 2011. Forecasting concentrations of air pollutants by logarithm support vector regression with immune algorithms. Appl. Math. Comput. 217, 5318–5327.
Liu, H., Motoda, H., 2012. Feature Selection for Knowledge Discovery and Data Mining. Springer Science & Business Media.
Liu, J., et al., 2019. Satellite-based PM2. 5 estimation directly from reflectance at the top of the atmosphere using a machine learning algorithm. Atmos. Environ. 208, 113–122.
Logan, W., 1953. Mortality in the London Fog Incident, 1952. pp. 336–338 Lancet.
MacKay, D.J., 1992. Bayesian interpolation. Neural Comput. 4, 415–447.
Maione, C., et al., 2018. Establishing chemical profiling for ecstasy tablets based on trace element levels and support vector machine. Neural Comput. Appl. 30, 947–955.
Matthias, V., et al., 2010. The contribution of ship emissions to air pollution in the North Sea regions. Environ. Pollut. 158, 2241–2250.
Melgani, F., Bruzzone, L., 2004. Classification of hyperspectral remote sensing images with support vector machines. IEEE Trans. Geosci. Remote Sens. 42, 1778–1790.
Mohamad, N., et al., 2016. Source apportionment and health risk assessment of PM10 in a naturally ventilated school in a tropical environment. Ecotoxicol. Environ. Saf. 124, 351–362.
Ning, X., et al., 2019. Ambient PM2. 5 causes lung injuries and coupled energy metabolic disorder. Ecotoxicol. Environ. Saf. 170, 620–626.
Niu, L., et al., 2015. Highly time- and size-resolved fingerprint analysis and risk assessment of airborne elements in a megacity in the Yangtze River Delta, China. Chemosphere 119, 112–121.
Qiu, Y.-N., et al., 2019. PM2. 5 induces liver fibrosis via triggering ROS-mediated mitophagy. Ecotoxicol. Environ. Saf. 167, 178–187.
Querol, X., et al., 2001. PM10 and PM2. 5 source apportionment in the Barcelona Metropolitan area, Catalonia, Spain. Atmos. Environ. 35, 6407–6419.
Reche, C., et al., 2012. A multidisciplinary approach to characterise exposure risk and toxicological effects of PM10 and PM2. 5 samples in urban environments. Ecotoxicol. Environ. Saf. 78, 327–335.
Ruan, J., et al., 2017. Improving the efficiency of dissolved oxygen control using an on-line control system based on a genetic algorithm evolving FWNN software sensor. J. Environ. Manag. 187, 550–559.
Shou, Y., et al., 2019. A review of the possible associations between ambient PM2. 5 exposures and the development of Alzheimer's disease. Ecotoxicol. Environ. Saf. 174, 344–352.
Sousa, S.I.V., et al., 2007. Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. Environ. Model. Softw 22, 97–103.
Sprague, B., et al., 2014. Design, synthesis and experimental validation of novel potential chemopreventive agents using random forest and support vector machine binary classifiers. J. Comput. Aided Mol. Des. 28, 631–646.
Srivastava, N., et al., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958.
Sun, W., Sun, J., 2017. Daily PM2.5 concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm. J. Environ. Manag. 188, 144–152.
Sun, W., et al., 2013. Prediction of 24-hour-average PM2.5 concentrations using a hidden Markov model with different emission distributions in Northern California. Sci. Total Environ. 443, 93–103.
Tian, Y., et al., 2018. Spatial, seasonal and diurnal patterns in physicochemical characteristics and sources of PM2.5 in both inland and coastal regions within a megacity in China. J. Hazard Mater. 342, 139–149.
Tsai, S.-S., Yang, C.-Y., 2014. Fine particulate air pollution and hospital admissions for pneumonia in a subtropical city: taipei, taiwan. J. Toxicol. Environ. Health, Part A. 77, 192–201.
Tsai, Y.I., 2005. Atmospheric visibility trends in an urban area in Taiwan 1961–2003. Atmos. Environ. 39, 5555–5567.
Tsai, Y.I., Chen, C.-L., 2006. Characterization of Asian dust storm and non-Asian dust storm PM2.5 aerosol in southern Taiwan. Atmos. Environ. 40, 4734–4750.
Tsai, Y.I., Cheng, M.T., 1999. Visibility and aerosol chemical compositions near the coastal area in Central Taiwan. Sci. Total Environ. 231, 37–51.
Tu, J.V., 1996. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J. Clin. Epidemiol. 49, 1225–1231.
Vapnik, V.N., 1999. An overview of statistical learning theory. IEEE Trans. Neural Netw. 10, 988–999.
Vlachogianni, A., et al., 2011. Evaluation of a multiple regression model for the forecasting of the concentrations of NOx and PM10 in Athens and Helsinki. Sci. Total Environ. 409, 1559–1571.
Vu, B., et al., 2012. Mutagenicity assessment of aerosols in emissions from wood combustion in Portugal. Environ. Pollut. 166, 172–181.

Wang, S., et al., 2017. A study of characteristics and origins of haze pollution in Zhengzhou, China, based on observations and hybrid receptor models. Aerosol and Air Quality Research 17, 513–528.
Wang, X., Sun, W., 2019. Meteorological parameters and gaseous pollutant concentrations as predictors of daily continuous PM2. 5 concentrations using deep neural network in Beijing–Tianjin–Hebei, China. Atmos. Environ. 211, 128–137.
Wang, Y., et al., 2019. Identification of mRNA-miRNA crosstalk in human endothelial cells after exposure of PM2. 5 through integrative transcriptome analysis. Ecotoxicol. Environ. Saf. 169, 863–873.
Wang, Z.-X., Hao, P., 2016. An improved grey multivariable model for predicting industrial energy consumption in China. Appl. Math. Model. 40, 5745–5758.
Wu, C.-D., et al., 2018. A hybrid kriging/land-use regression model to assess PM2. 5 spatial-temporal variability. Sci. Total Environ. 645, 1456–1464.
Wu, C.-F., et al., 2017. Regulation of fine particulate matter (PM 2.5) in the pacific rim: perspectives from the APRU global health program. Air Quality, Atmosphere Health 10, 1039–1049.
Wu, P.-C., et al., 2004. Increased levels of ambient fungal spores in Taiwan are associated with dust events from China. Atmos. Environ. 38, 4879–4886.
Yang, C.-Y., et al., 2004. Relationship between air pollution and daily mortality in a subtropical city: taipei, Taiwan. Environ. Int. 30, 519–523.
Yang, F., et al., 2014. Particle size-specific distributions and preliminary exposure assessments of organophosphate flame retardants in office air particulate matter. Environ. Sci. Technol. 48, 63–70.
Yu, W., et al., 2017. Improving satellite-based PM 2.5 estimates in China using Gaussian processes modeling in a Bayesian hierarchical setting. Sci. Rep. 7, 7048.
Zhan, Y., et al., 2017. Spatiotemporal prediction of continuous daily PM2. 5 concentrations across China using a spatially explicit machine learning algorithm. Atmos. Environ. 155, 129–139.
Zhang, J., et al., 2019. The acute health effects of ozone and PM2. 5 on daily cardiovascular disease mortality: a multi-center time series study in China. Ecotoxicol. Environ. Saf. 174, 218–223.
Zheng, L., et al., 2019. Seasonal and spatial variations of PM10-bounded PAHs in a coal mining city, China: distributions, sources, and health risks. Ecotoxicol. Environ. Saf. 169, 470–478.
Zhou, Q., et al., 2014. A hybrid model for PM2.5 forecasting based on ensemble empirical mode decomposition and a general regression neural network. Sci. Total Environ. 496, 264–274.
Zhu, S., et al., 2018. Optimal-combined model for air quality index forecasting: 5 cities in North China. Environ. Pollut. 243, 842–850.