Air Pollution Modelling from Meteorological Parameters Using Artificial Neural Network

Sateesh N. Hosamane[1(✉)] and G. P. Desai[2]

[1] Department of Chemical Engineering, K. L. E. DR. M. S. Sheshgiri College of Engineering and Technology, Belagavi 590 008, Karnataka, India
satishosamane@gmail.com

[2] Department of Chemical Engineering, Bapuji Institute of Technology, Davanagere 577005, Karnataka, India
desai_gp@yahoo.com

Abstract. The aim of this study is to develop neural network air quality prediction model for PM_{10} (particle whose diameter is less the 10 μm), NO_2 and SO_2. A multilayer neural network model with a hidden recurrent layer is used to predict pollutant concentrations at four monitoring sites in Belagavi city of Karnataka State, India. The Levenberg Marquardt algorithm is used to train the network. A combination of input variables were investigated taking into the predictability of meteorological input variables and the study of model performance. The meteorological variables air temperature, wind speed, wind direction, rainfall and relative humidity were considered as input variables for this study. The results show very good agreement between measured and predicted pollutant concentrations. The performance of the developed model was assessed through performance index. The models developed have good prediction performance (>85%) for all the pollutants. The proposed models were predicted pollutant concentration with relatively good accuracy and outputs were proven to be satisfactory by measuring of the goodness of fit and by mean absolute percentage error.

Keywords: ANN · Air pollution · Modelling · Prediction

1 Introduction

Air pollution is a serious environmental problem across major cities in India. The problem of air pollution has significant health impacts on human and the environment [1, 2]. Atmospheric pollution sourced by industrial activities and congestion of roads and traffic is of principal concern. Air pollution concentration is mainly due to various combination of pollutant and their physiochemical interactions or processes with other components in the atmosphere, properties of earth surface and geometry [3]. PM_{10}, SO_2 and NO_2 are the three main pollutants responsible for the degradation of the ambient air quality. These pollutants exert a wide range of impacts on biological, physical, and economic systems, especially, effects on plant and human health are of particular

D. J. Hemanth and S. Smys (eds.), *Computational Vision and Bio Inspired Computing*,
Lecture Notes in Computational Vision and Biomechanics 28,
https://doi.org/10.1007/978-3-319-71767-8_39

concern. Worldwide there are more deaths due to poor air quality than from automobile accidents [4]. Air pollution is a serious problem in urban areas, expose to higher air pollutant concentrations for longer period may increase the risk of asthma, respiratory and cardiovascular systems, cancer and mortality. Therefore, there is an urgent need to development an efficient forecasting system to provide air quality information to the general public. In recent years statistical methods have been used for air pollution predictive models which are not capable of predicting short term pollution levels. Regression modelling is most popular statistical approach has been used to develop air quality predictive models in number of studies [5]. Linear regression models have good predictability with linear process. They will underperform when we try to model nonlinear processes. Therefore, artificial neural networks (ANN) approach is the best as compared with statistical linear methods, especially where the problem being analyzed includes nonlinear behavior [6]. ANN models have been used for the forecasting a wide range of pollutants and their concentrations at various time scales, with very good results [7, 8]. The modelling tools are widely used in many scientific fields, especially in environmental sciences and have been widely applied for modelling air pollutant concentrations with the aim to forecast them. ANN modelling is the best tool for nonlinear relationships. Their performance capability is superior when compared to other statistical methods [9, 10]. The literature showed that fairly good estimates can be achieved by different models developed using ANN. The most common structure of the neural network is the "feed forward" where the data flow from input to output units is strictly feed forward and are able to find and identify complex patterns in datasets which may not be well described by a simple mathematical formula or a set of known processes [11]. The neural network approach is applied for highly complex pattern recognition and to solve the problem in presence of noisy dataset [12, 13]. ANN models performed better when they combined with traditional deterministic modelling approach [14]. In this paper, the back propagation algorithms are used to model and predict the daily average concentrations of PM_{10}, NO_2 and SO_2 using meteorological variables (Input).

1.1 Artificial Neural Network (ANN)

ANNs are computing systems motivated by biological models, and made up of a number of easy and highly interconnected processing components, which process information by its dynamic state response to external inputs. The processing components called neurons are organized into interconnected layers [14–16]. Multilayer perceptron (MLP) was introduced by Rumelhart in 1986. It is most widely used and composed of three layers of neurons. Several neurons are organised into input layer, hidden layer and the output layer. The input layers take the value from the model input and serves to pass the values to the hidden layer as shown in Fig. 1. In this study, feed forward ANN was used to predict air pollutant concentrations based on different meteorological variables. The number of hidden layers were selected by using trial and error procedure varying for 8–12 hidden layers in the network structure was examined in order to check the effectiveness of network predictability.

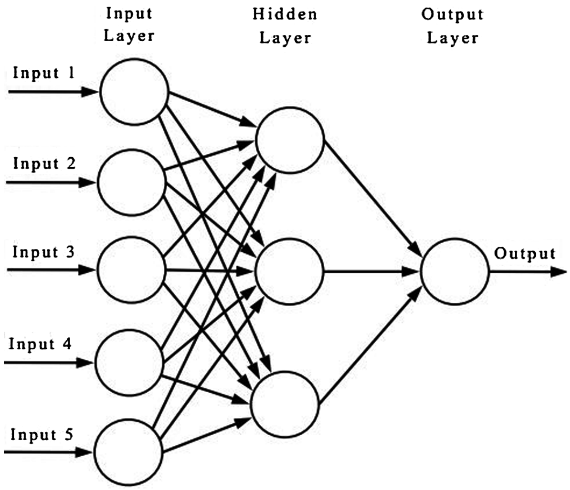

Fig. 1 Architecture of multilayer neural network

2 Study Area

Figure 2 shows satellite view of Belagavi city highlighted with four monitoring stations, is located at 15.87 °N 74.5 °E, with an average elevation of 751 m. The city is in the north western parts of Karnataka state and is along the border of two states, Goa and Maharashtra. Belagavi is a district head quarter and officially known as second capital of Karnataka State, India.

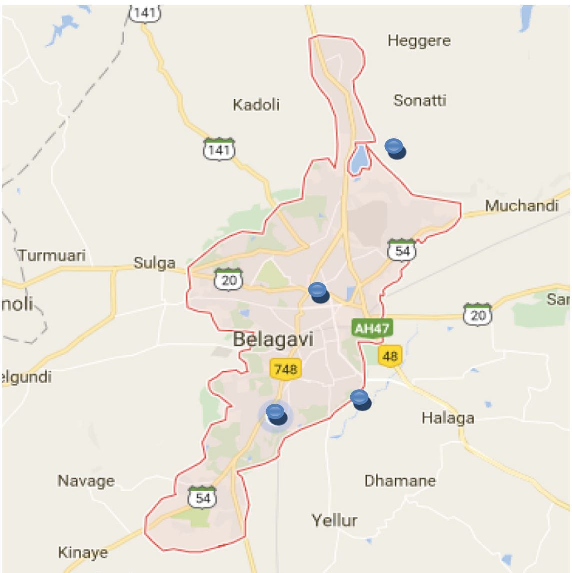

Fig. 2 Satellite picture of Belagavi city showing four monitoring sites

Belagavi is known for its foundry Hub. More than 200 foundries are producing automotive and industrial castings of ferrous base and supporting ancillaries and one of the largest alumina producing company is located at Belagavi city. Vehicle population and traffic, industrial emissions are the major sources of pollution in the Belagavi atmosphere, a problem that has been annoyed by the drastic increase in the number of mobile sources from last five years. Therefore, there is an urgent need for the assessment and evaluation of air quality in Belagavi city. It has been found that air pollutant concentrations of PM_{10}, SO_2, and NO_2 distribution is strongly affected by meteorological factors.

2.1 Data Sets

The meteorological data temperature (in °C), wind speed (in mps), wind direction (in degrees), relative humidity (in %) and rain (in mm) were collected from meteorological station located at Sambra, Belagavi city. Table 1 shows the mean, standard deviation, minimum and maximum values of meteorological parameters and pollutant concentrations for the year 2011–2013. The mean annual temperature is around 24.65 °C, annual relative humidity is between 17.15 and 99.07% and prevailing wind direction NNE to NNW and annual wind speed is between 2 and 13.73 mps and heavy rain occurs in the period of June to September is between 0 and 9.23 mm.

Table 1 Daily average means, standard deviation, minimum and maximum values of Meteorological parameters and pollutant concentrations of PM_{10}, SO_2 and NO_2

Variable	Mean		Standard deviation		Minimum			Maximum
Temperature, °C	24.65		2.57		20.12			34.09
WS, mps	6.58		2.42		2.00			13.73
WD, degrees	195.09		82.52		22.05			337.51
Rain, mm	0.42		0.88		0.00			9.23
Humidity, %	66.04		20.93		17.51			99.07
Autonagar					*Railway Station*			
	Mean	SD	Min	Max	Mean	SD	Min	Max
PM_{10}, μg/m^3	63.07	34.88	12.35	245.85	88.71	34.46	15.03	208.95
SO_2, μg/m^3	5.98	2.36	0.56	14.87	7.71	2.44	1.98	17.57
NO_2, μg/m^3	20.98	11.98	2.43	77.56	26.58	9.65	7.16	61.26
Vadgaon					*Udyambag*			
	Mean	SD	Min	Max	Mean	SD	Min	Max
PM_{10}, μg/m^3	46.75	17.41	10.15	101.23	53.78	18.25	9.87	152.36
SO_2, μg/m^3	3.97	1.58	0.98	10.56	10.11	4.87	1.06	28.36
NO_2, μg/m^3	12.16	5.33	1.66	36.91	15.32	7.02	3.02	53.30

The site characteristics are different resulting industrial (Autonagar and Udyambag), traffic (Railway station) and residential (Vadgaon). Air quality monitoring sites were classified as industrial, traffic and residential sites and the air pollution monitoring was carried for the period of 2011–2013. Respirable dust sampler RDS (Envirotech APM 460 NL) was used to monitor and measure the concentration of PM_{10} ($\mu g/m^3$) and Gaseous pollutant sampler (Environtech APM 433) was used to measure NO_2 ($\mu g/m^3$) and SO_2 ($\mu g/m^3$) concentrations using suitable reagents. The data sets of meteorological parameters and pollutant concentrations are used on their daily average for the period of 2011–2013.

We had problems in the data sets mainly outlier and missing data. The outlier was due to the malfunction of instrument or incorrect measurement of the pollutant. The outliers are maximum or minimum values of data. They are analysed with care, because they cause more deviation in the model development and prediction. Missing data was due to instrument calibration or malfunctions and this problem was very limited (2%), these gaps are filled by linear interpolation method [16]. In order to support the neural network to efficiently handle a data, all the input variables are normalised to the range (0, 1) by Eq. 1.

$$X_n = \frac{(X_i - X_{\min})}{(X_{\max} - X_{\min})} \tag{1}$$

X_n is the normalised data, X_i actual measured data, $X_{\min}$ minimum value of the measured data, $X_{\max}$ maximum value of the measured data.

2.2 ANN Modelling

The ANN models were developed by using the MATLAB R2015a software from Mathworks group Inc. The feed forward Back propagation (BP) multilayer perceptron (MLP) network model was used for the present study. The BP algorithm is used to train a given feed-forward multilayer neural network for a known set of input patterns with known classifications. The BP algorithm is based on Widrow-Hoff delta learning rule which is based on weight adjustment through mean square error of the output to the sample input [17]. Therefore, BP networks are the simplest and most widely used network models [18]. The meteorological variables (inputs) and pollutant concentrations (outputs) are divided into training, validation and testing subsets. One third data was used for validation, one third data was used for testing and two third of the data was used for the training set. The neuron number in hidden layer has the significant importance for the model development and its accuracy and performance. Another important step in model development is determination of activation function. The most widely used activation functions are liner, sigmoid and hyperbolic tangent. The optimization of one hidden layer was conducting by several tests with various network structures, optimised networks have selected based on lower prediction error and smaller convergence times. Present study applied back propagation network with three layers to predict air pollutant concentrations. Selected network structure was used and trained after definition of subset with *Levenberg-Marquardt* optimization (*trainlm*) in hidden layer (nodes of 10, 12, 18), log-sigmoid (*logsig*) transfer function is used in output layer [19, 20].

2.3 Evaluation of Model Performance

Evaluation of the performance of developed model is very important and is important to evaluate forecast accuracy. We have selected statistical indicator to describe goodness of the estimates. The accuracy of model was determined by considering how well a model performed with new data which are not used in model fitting. The model performance was checked using Mean Absolute Error (MAE), and Root Mean-Square Error (RMSE) and correlation Coefficient (R) are given by Eqs. (2)–(6).

$$MSE = \frac{\sum_{i=1}^{n} (Y_i - X_i)^2}{n} \tag{2}$$

$$MAE = \frac{\sum_{i=1}^{n} |Y_i - X_i|}{n} \tag{3}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - X_i)^2}{n}} \tag{4}$$

$$MAPE = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - X_i)^2}{n}} \tag{5}$$

$$R = \frac{\sum_{i=1}^{n} \left(Y_i - \overline{Y_i}\right)\left(X_i - \overline{X_i}\right)}{\left\{\left[\sum_{i=1}^{n} \left(X_i - \overline{X_i}\right)^2\right]\left[\sum_{i=1}^{n} \left(Y_i - \overline{Y_i}\right)^2\right]\right\}^{1/2}} \tag{6}$$

where, n is the number of data, Y_i is the modeled pollutant concentration, X_i is the observed concentration. Zero Error indicates that all the modeled concentrations of various pollutants computed by ANN models were perfectly match the observed concentrations.

3 Results and Discussions

Overall, the annual average concentrations of PM_{10} were above the national ambient air quality standard (NAAQS) at Autonagar and Railway station and is almost at alarming stage at Udyambag and Vadgaon. PM_{10} concentrations are mainly due to construction activity, industrial activity and bad road conditions. The annual average NO_2 and SO_2 concentrations were below NAAQS standards. The concentrations of NO_2 are higher on traffic sites (Railway station) and industrial area (Autonagar and Udyambag) and it confirms industries and traffic as important sources of NO_2 in Belagavi city. SO_2 concentration is very low (negligible) and is mainly due to diesel vehicles and old vehicles, industrial burning and commercial burning of various fuel oils.

The optimisation of a neural network is most important objective to developed ANN based models [21]. The process of optimization plays an important role in the selection and performance of the network. Hence, an optimisation was carried out with number of neurons and MSE [19]. Then, the multilayer layer neural network was evaluated using BP algorithm with 10, 12 and 18 nodes in the hidden layer. With increasing in neuron number, the network gave several local minimum values and different MSE values were obtained for the training set. Increasing neuron number to more than 20 gave unrealistic results for all the pollutants.

The various performance indicators were used to determine to measure the goodness of the fit and the results of ANN model are summarized in Table 2. The best performing ANN network was *trainlm* in the hidden layer. The results shows excellent performance for the developed models for PM_{10} for all four monitoring sites are optimised with nodes of 10, 10, 18 and 12 according to values of MAPE was found to be 7.2, 6.2, 10.8, and 7.7%. Good results were obtained for NO_2 with nodes of 18, 12, 10, 10 with values of MAPE are 12.1, 3.5, 8.6, 8.7%. Similarly, good performance by the developed network of SO_2 for all monitoring sites using nodes of 12, 12, 18 and 10 with MAPE of 2.0, 7.1, 3.6 and 8.6%. The smallest MSE was obtained for *trainlm* function and the mean absolute percentage error was used to select the models. The following models have performed best based on the change in number of hidden nodes and keeping the *logsig* transfer function. The training stopped after 8, 9 and 9 iterations for PM_{10}, The training stopped after 10, 11 and 9 iterations for SO_2, for NO_2, training stopped after 6, 7 and 8 iterations.

Table 2 Performance indices for the testing data for four monitoring sites

PM_{10}—Autonagar						PM_{10}—Railway station				
Nodes	MAD	MSE	RMSE	MAPE	R	MAD	MSE	RMSE	MAPE	R
10	47.56	71.88	6.757	0.072	0.9592	88.18	28.71	4.212	0.061	0.8309
12	64.56	74.18	7.199	0.200	0.9072	88.70	94.39	7.691	0.114	0.8467
18	65.71	41.91	5.587	0.163	0.9447	90.40	141.6	9.613	0.149	0.8550
PM_{10}—Udyambag						PM_{10}—Vadgaon				
Nodes	MAD	MSE	RMSE	MAPE	R	MAD	MSE	RMSR	MAPE	R
10	48.19	27.90	4.384	0.130	0.8420	52.11	29.15	4.409	0.093	0.8863
12	46.22	28.93	4.367	0.115	0.8420	50.79	26.60	4.152	0.077	0.9015
18	47.09	21.92	3.851	0.108	0.8511	58.02	29.09	4.594	0.123	0.8982
NO_2—Autonagar						NO_2—Railway station				
Nodes	MAD	MSE	RMSE	MAPE	R	MAD	MSE	RMSE	MAPE	R
10	3.911	5.06	2.075	0.341	0.8389	5.462	5.101	2.155	0.280	0.8785
12	4.240	3.58	1.745	0.287	0.8632	5.681	4.125	1.937	0.252	0.8899
18	5.186	0.981	0.819	0.121	0.8982	7.716	0.061	0.191	0.035	0.8925
NO_2—Udyambag						NO_2—Vadgaon				
Nodes	MAD	MSE	RMSE	MAPE	R	MAD	MSE	RMSE	MAPE	R
10	4.576	0.3824	0.603	0.223	0.8721	9.589	1.221	0.847	0.087	0.8781
12	4.087	0.0534	0.1978	0.086	0.9249	10.68	1.92	1.194	0.190	0.8927
18	4.517	0.3412	0.549	0.216	0.8828	10.54	1.084	0.896	0.144	0.9320

(continued)

Table 2 *(continued)*

SO_2—Autonagar						SO_2—Railway station				
Nodes	MAD	MSE	RMSE	MAPE	R	MAD	MSE	RMSE	MAPE	R
10	20.14	5.05	1.432	0.063	0.9063	29.58	14.84	3.434	0.174	0.8673
12	20.59	0.59	0.485	0.020	0.9690	25.15	6.15	1.939	0.071	0.9014
18	22.10	1.44	1.200	0.072	0.9877	22.00	6.00	1.501	0.080	0.8847
SO_2—Udyambag						SO_2—Vadgaon				
Nodes	MAD	MSE	RMSE	MAPE	R	MAD	MSE	RMSE	MAPE	R
10	13.39	1.021	0.912	0.100	0.9348	14.58	4.12	1.363	0.086	0.8531
12	12.70	0.204	0.354	0.036	0.9568	15.18	3.60	1.340	0.099	0.8884
18	13.60	0.411	0.324	0.078	0.9435	15.78	3.296	1.399	0.116	0.8916

Figure 3 shows observed and predicted PM_{10}, NO_2 and SO_2 concentrations using developed ANN models. It demonstrates that these models can estimate air pollutant concentration for the given set with an accuracy of approximately 90%. As discussed earlier the used data set was divided into two subsets, the first set was included data from the period of 2011–2012 and 2013 data was totally unknown data for the model was used to evaluate the forecasting ability of the developed model is shown in Fig. 4. According to Fig. 4, it seems that the prediction of all developed models is in a very satisfactory level ($p < 0.05$). The performance of developed models for training data set and performance in terms R are shown in Fig. 4. The results indicate that developed ANN models predicated the PM10 with good accuracy of 95.92, 85.50, 85.11 and 90.15%, NO_2 with 89.82, 89.27, 92.49 and 93.20%, for SO_2 with excellent accuracy of 98.77, 88.47, 95.68 and 89.16% for the monitoring sites Autonagar, Railway station, Udyambag and Vadgaon.

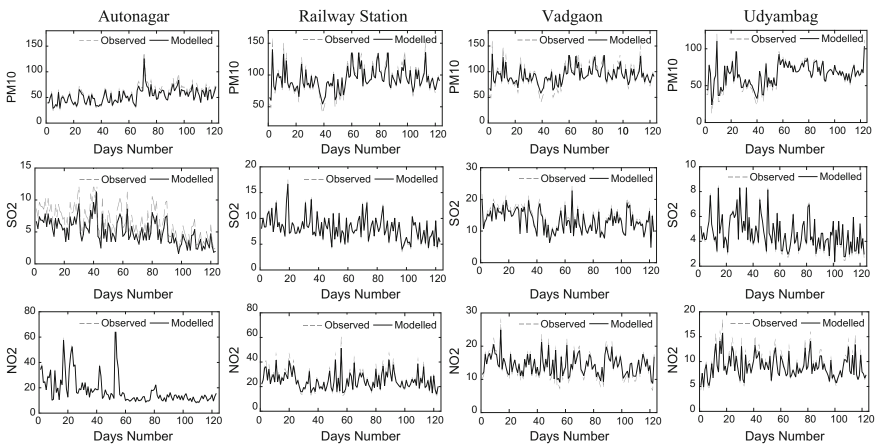

Fig. 3 Measured and predicted concentrations of PM_{10}, SO_2 and NO_2 for training data

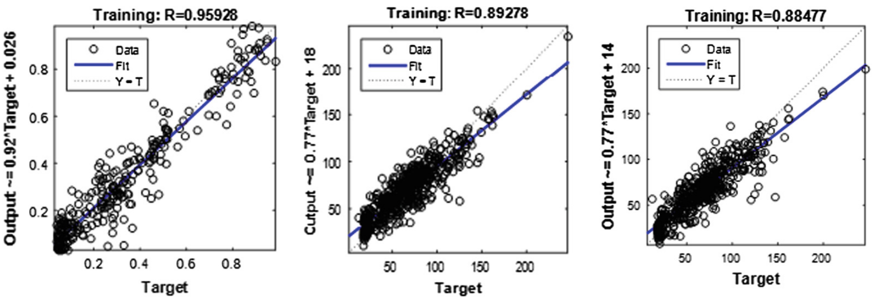

Fig. 4 ANN outputof training data set for PM_{10} (Autonagar), NO_2 (Vadgaon) and SO_2 (Railway station)

4 Conclusion

The optimization study was done for better ANN mode with different structures in terms of hidden layer and number of nodes. Three layer model used for the present study showed a precise and effective prediction of air pollutant concentrations. The *trainlm* has given good prediction with relatively good accuracy. For the air pollution predictive the ANN models have faster predictive power and are viable as compared to other statistical models. The performance of the model in all four locations was satisfactory and is able to efficiently simulate the atmospheric time-series pollutant concentrations. ANN models can be considered as appropriate for operational usage in urban air quality management. The models discussed in this study are easily implemented and are helpful to the local authorities in providing the information to the general public and to protect the health of people by implementing appropriate controlling measures.

References

1. Dimitriou, K., Paschalidou, A.K., Kassomenos, P.A.: Assessing air quality with regards to its effect on human health in the European Union through air quality indices. Ecol. Ind. **27**, 108–115 (2013)
2. Pope III, C.A., Burnett, R.T., Thun, M.J., Calle, E.E., Krewski, D., Ito, K., Thurston, G.D.: Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. JAMA **287**(9), 1132–1141 (2002)
3. Karatzas, K.D., Kaltsatos, S.: Air pollution modelling with the aid of computational intelligence methods in Thessaloniki. Greece. Simul. Model. Pract. Theor. **15**(10), 1310–1319 (2007)
4. Deleawe, S., Kusznir, J., Lamb, B., Cook, D.J.: Predicting air quality in smart environments. J. Ambient Intell. Smart Environ. **2**(2), 145–154 (2010)
5. Gardner, M.W., Dorling, S.R.: Neural network modelling and prediction of hourly NO_x and NO_2 concentrations in urban air in London. Atmos. Environ. **33**(5), 709–719 (1999)
6. Viotti, P., Liuti, G., Di Genova, P.: Atmospheric urban pollution: applications of an artificial neural network (ANN) to the city of Perugia. Ecol. Model. **148**(1), 27–46 (2002)

7. Karaca, F., Alagha, O., Ertürk, F.: Application of inductive learning: air pollution forecast in Istanbul. Turkey. Intell. Autom. Soft Comput. **11**(4), 207–216 (2005)
8. Athanasiadis, I.N., Karatzas, K., Mitkas, P.: Contemporary air quality forecasting methods: a comparative analysis between classification algorithms and statistical methods. In: Fifth International Conference on Urban Air Quality Measurement, Modelling and Management, Valencia, Spain (2005)
9. Kolehmainen, M., Martikainen, H., Ruuskanen, J.: Neural neworks and periodic components used in air quality forecasting. Atmos. Environ. **35**(5), 815–825 (2001)
10. Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., Cawley, G.: Extensive evaluation of neural network models for the prediction of NO_2 and PM_{10} concentrations, compared with a deterministic modelling system and measurements in central Helsinki. Atmos. Environ. **37**(32), 4539–4550 (2003)
11. Rumelhart, E., Hinton, J., Williams, R.: Learning internal representations by error propagation, in parallel distributed processing: exploration in the microstructure of cognition, vol. 1. MIT press, Cambridge (1986)
12. Hertz, J.A., Krogh, A.S., Palmer, R.G.: Introduction to the theory of neural computation. Addison Wesley, Canada (1995)
13. Bishop, A.: Neural networks for pattern recognition. Oxford University Press, UK (1995)
14. Fausett, L.: Neural Networks: Architectures, Algorithms, and Applications. Prentice-Hall Inc., New Jersey (1994)
15. Gardner, M.W., Dorling, S.R.: Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmos. Environ. **32**(14), 2627–2636 (1998)
16. Kandasamy, S., Baret, F., Verger, A., Neveux, P., Weiss, M.: A comparison of methods for smoothing and gap filling time series of remote sensing observations application to MODIS LAI products. Biogeosciences **10**(6), 4055–4071 (2013)
17. http://wwwold.ece.utep.edu/research/webfuzzy/docs/kk-thesis/kk-thesis-html/node22.html
18. Niska, H., Hiltunen, T., Karppinen, A., Ruuskanen, J., Kolehmainen, M.: Evolving the neural network model for forecasting air pollution time series. Eng. Appl. Artif. Intell. **17**(2), 159–167 (2004)
19. Velasquez, G.: A Distributed approach to a neural network simulation program. Master's thesis, The University of Texas at El Paso, El Paso (1998)
20. Cai, M., Yin, Y., Xie, M.: Prediction of hourly air pollutant concentrations near urban arterials using artificial neural network approach. Transp. Res. Part D Transp. Environ. **14**(1), 32–41 (2009)
21. Akkoyunlu, A., Yetilmezsoy, K., Erturk, F., Oztemel, E.: A neural network-based approach for the prediction of urban SO_2 concentrations in the Istanbul metropolitan area. Int. J. Environ. Pollut. **40**(4), 301–321 (2010)