Advances in Environmental Research, *Vol. 4, No. 2* (*2015*) *83-104*
DOI: http://dx.doi.org/10.12989/aer.2015.4.2.083

Multicity Seasonal Air Quality Index Forecasting using Soft Computing Techniques

Shruti S. Tikhe [1a], K.C. Khare [*2] and S.N. Londhe [3b]

[1] *Department of Civil Engineering*, *Sinhgad College of Engineering*, *Pune Maharashtra*, *India - 411041*
[2] *Department of Civil Engineering Symbiosis Institute of Technology*, *Pune Maharashtra*, *India - 412115*
[3] *Department of Civil Engineering*, *Vishwakarma Institute of Information Technology Pune Maharashtra*, *India - 411048*

(*Received December 17, 2014, Revised May 28, 2015, Accepted June 08, 2015*)

Abstract. Air Quality Index (AQI) is a pointer to broadcast short term air quality. This paper presents one day ahead AQI forecasting on seasonal basis for three major cities in Maharashtra State, India by using Artificial Neural Networks (ANN) and Genetic Programming (GP). The meteorological observations & previous AQI from 2005-2008 are used to predict next day's AQI. It was observed that GP captures the phenomenon better than ANN and could also follow the peak values better than ANN. The overall performance of GP seems better as compared to ANN. Stochastic nature of the input parameters and the possibility of auto-correlation might have introduced time lag and subsequent errors in predictions. Spectral Analysis (SA) was used for characterization of the error introduced. Correlational dependency (serial dependency) was calculated for all 24 models prepared on seasonal basis. Particular lags (*k*) in all the models were removed by differencing the series, that is converting each *i*′th element of the series into its difference from the (*i*-*k*)″th element. New time series is generated for all seasonal models in synchronization with the original time line & evaluated using ANN and GP. The statistical analysis and comparison of GP and ANN models has been done. We have proposed a promising approach of use of GP coupled with SA for real time prediction of seasonal multicity AQI.

Keywords: Air Quality Index; ANN; GP; spectral analysis

1. Introduction

Deteriorating urban air quality is a fundamental problem all over the world which is further aggravated by pollution episodes. Persistent exposure to urban air pollution affects human health. Hence it is necessary to develop a system which would inform public about the existing and future air quality.

The Air Quality Index (AQI) is a scale designed to understand what the air quality is. It tells how clean or polluted the air is and what associated health effects might be a concern for human

∗Corresponding author, Professor, E-mail: kanchan.khare@sitpune.edu.in
[a] Research Scholar
[b] Professor

lives. The AQI focuses on health effects one may experience within a few hours or days after breathing polluted air (Mohan and Anurag 2007). It serves as a warning to make decisions to protect ones health by limiting short term exposure to air pollution and adjusting ones activities during increased levels of air pollution. The AQI varies from 0 to 500. The higher the AQI value, the greater the level of air pollution and the greater are the health concerns. An AQI value of 100 generally corresponds to the national air quality standard for the pollutant, which is the level, EPA (Environment Protection Authority) has set to protect public health in India. AQI values below 100 are generally thought of as satisfactory. When AQI values are above 100, air quality is considered to be unhealthy at first for certain sensitive groups of people, then for everyone as the AQI values get higher as detailed in Table 1.

Advances in mathematical models to describe the formation, emission, transport and disappearance of air pollutants have led to a greater understanding of the dynamics of these pollutants. But complex models are generally information hungry. A lot of data is necessary for their application to have sufficient certainty that the results will have technical or scientific value (Russell and Dennis 2000). These deterministic models require much information that is not always possible to obtain; the data available have not always resulted in successful outcomes upon application of the model (Roth 1999) or the cost of obtaining reliable data can be prohibitive (Pun *et al*. 2000). This is more applicable to developing countries.

There are other methods requiring less information that can be used to study air pollution in some areas. These methods generally make use of statistical techniques such as regression or other data-fitting methods using numerical techniques to establish the respective relationships between the various physicochemical parameters and variable of interest based on routinely-measured historical data.

The main objectives of these methods include investigating and assessing trends in air quality, making environmental forecasts and increasing scientific understanding of the mechanisms that govern air quality (Thompson *et al*. 2001). Among the techniques being examined to relate air quality in a given area to measured physical and chemical parameters, the three that have been used are:

(1) Multivariate regression (Hubbard and Cobourne 1998, Comrie and Diem 1999, Davis and Speakman 1999, Draxler 2000, Gardner and Dorling 2000),
(2) Artificial neural networks (ANN) (Perez and Reyes 2006, Brunelli *et al*. 2007, Thomas and Jacko 2007, Grivas and Chaloulakou 2005, Gardner and Dorling 2000) and
(3) Time series and spectral analysis (Raga and Le Moyne 1999, Chen *et al*. 1998, Milanchus *et al*. 1998, Salcedo *et al*. 1999, Sebald *et al*. 2000).

Fourier Transforms would have been a common way to quantify the periodicities in time series data. However because of irregularities or missing data the Fast Fourier Transform (FFT) methods cannot be successfully applied to environmental studies and air quality analysis (Dilmanghanani 2007). Due to this fact spectral analysis has not been a common method in literature. Most documented air quality examinations are based on statistical methods such as means and variance (Dilmanghanani 2007).

One more (soft computing) technique of Fuzzy logic is also found to be used for air quality modeling and forecasting.

Both ANN and Fuzzy logic resemble either biological processes or mimic the ability of the human brain to employ effectively the modes of reasoning that are approximate rather than exact

and result into a quicker solution. Conventional approach of hard computing techniques require precisely stated analytical model and a lot of computational time. Hard computing techniques are based on the principles of precision, certainty and rigor. Many real life problems cannot be solved using hard computing as they do not lend themselves to precise solutions. In this situation, soft computing supersedes hard computing because they are tolerant of imprecision, uncertainty, partial truth and approximation to achieve tractability, robustness and low solution cost (Zadeh 1994). Major soft computing tools are Artificial Neural Networks, Fuzzy logic, Genetic Programming, Support Vector Machines etc.

In the recent years, AQI has been calculated as well as forecasted by many scientists all over the world. Rao *et al*. (2002) developed long term AQI for four major Indian cities including Mumbai, Delhi, Kolkata and Chennai and also the rapidly urbanizing area around Nagpur. AQI for the city of Kanpur was developed by Sharma *et al*. (2003) with an understanding of the relationship between seasonal effects and specific pollutant species comprising the AQI. Jiang *et al*. (2004) used MLP ANN model to develop AQI for Shanghai. Kyrkili *et al*. (2007) developed an aggregate air quality based on the combined effects of five criteria pollutants (CO, SO_2, NO_2, O_3 and PM_{10}) by taking into account European standards.

Van den Elshout *et al*. (2008) reviewed existing AQI and presented a proposal for common alternative. Mayer *et al*. (2008) evaluated long term AQI for Germany. Fuzzy inference systems are used by Hajek and Olej (2009) for modeling air quality index. Kumar and Goyal (2011) tried Principal Component Analysis for forecasting air quality of Delhi.AQI for Thessaloniki, one of the most polluted city of Greece was predicted by Kassomenos *et al*. (2012) and found that PM_{10} is the main contributor of the poor air quality. Wong *et al*. (2013) developed a health risk based AQI for Hong Kong.

Use of GP in the field of air quality prediction is relatively new.GP has been tried by Pires *et al*. (2011) as well as by Tikhe Shruti *et al*. (2013) in order to predict the pollutant concentrations a few time steps in advance .But there is no evidence of use of GP for AQI prediction. As far as air quality index prediction for major cities of India is concerned, Indian Institute of Tropical Meteorology has developed AQI forecasting numerical models under SAFAR (System of Air Quality Weather Forecasting and Research) project (http://safar.tropmet.res.in/) for Pune and Delhi. Both the models are operational and Mumbai, Chennai, Kolkata and Ahmedabad models are forthcoming.

In the present work seasonal AQI models for three major cities of Maharashtra are developed using soft computing techniques of Artificial Neural Networks (ANN) and Genetic Programming (GP).

2. Study area and data

Mumbai, Pune and Nashik (Fig. 1) are the three major cities of Maharashtra. They are at the top positions as far as pollution of the state is concerned (www.hoparoundindia.com). Mumbai is the capital city of Maharashtra state of India (http://en.wikipedia.org/wiki/Mumbai). Mumbai's climate can be best described as moderately hot with high level of humidity. The heavy fog enveloping the city during early mornings is an ominous indicator of Mumbai's deteriorating air quality. Owing to bustle of vehicles and construction, the pollution levels have surpassed the standard limits by two-fold (http://www.dnaindia.com).

Pune is one of the fastest developing metropolitan cities of India. It is considered to be the premier

industrial centre of the country. Rapid industrialization and construction activities have resulted into exponential vehicular population growth. This has worsened the air quality which is evident from increased pollutant concentrations for last few years.

Nashik is the third largest city of Maharashtra after Mumbai and Pune. Accelerating growth in the transport sector, booming construction industry and growing industrial sector are responsible for deteriorating air quality of the city, which has resulted into bad health impacts.

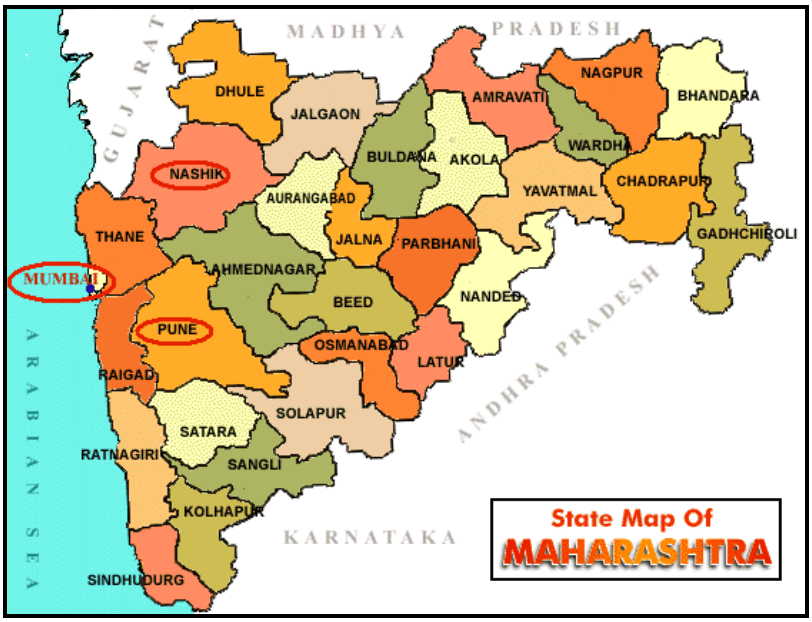

Fig. 1 Map of Maharashtra state of India (source maps of India.com)

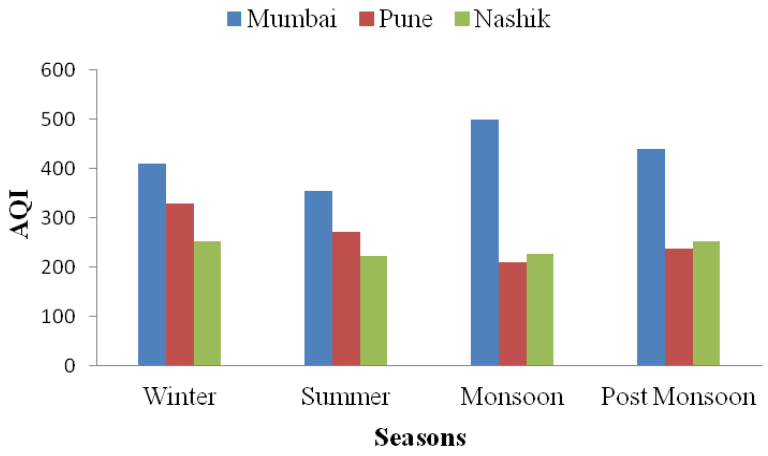

Fig. 2 Seasonal maximum AQI

Describing air quality of these three cities would give a broader picture of air quality of urban areas of Maharashtra. One fourth of the total population of Maharashtra lives in these three cities. Notifying one day ahead AQI would help to safeguard the health of one fourth of the total population of Maharashtra.

From the above graph (Fig. 2), it can be seen that the air quality of Mumbai is highly disturbed compared to Pune and Nashik. Maximum AQI recorded for Mumbai (500) in Monsoon, Pune (329) in Winter and Nashik (253) in Post Monsoon season in the study years (2005 to 2008). Responsible pollutant for highest AQI is RSPM (Respirable Suspended Particulate Matter) for all the three cities. RSPM have aerodynamic diameter less than or equal to 10 micrometers and are produced from combustion processes, vehicles and industrial sources. Higher concentrations of RSPM often results into respiratory and cardiovascular diseases and are reported to affect the sensitive groups.

Rainfall actually is a cleansing mechanism. Pollutants are washed away resulting the dropping of AQI. Still highest AQI for Mumbai is observed for monsoon season. The possible reason for highest AQI for Mumbai could be below average rainfall for the years 2006, 2007 and 2008 as observed from recoded data.(The average rainfall in Mumbai is 242 cm and the rainfall recorded is around 100 cm in these years). Winters in Pune are generally bad. Less circulation and more stagnant air masses result into lower rate of the chemical removal of gaseous pollutants and hence highest AQI.

Nashik is one of the important pilgrimage in Maharashtra. People from all over India visit Nashik especially during some specific festivals which happen to be in Post monsoon season resulting into bad air quality and highest AQI .Vehicular exhaust is one of the major source of RSPM as the vehicular polulation of Mumbai, Pune & Nashik is also very high compared to other cities of Maharashtra.

3. Materials and methods

Data used for the study consists of daily average values of surface meteorological parameters and pollutant concentrations recorded for the period of January 2005 to December 2008 by India Meteorological Department (IMD) (www.imdpune.gov.in) and Maharashtra Pollution Control Board (MPCB) (www.mpcb.gov.in) for Pune, Mumbai and Nasik respectively.

Meteorological parameters such as rainfall (RF) is measured by rainguages, temperature difference (TD) is recorded by thermometer, relative humidity (RH) is measured by hygrographs, station level pressure & vapour pressure (SLP & VP) are measured by barometer ,wind speed (WS) is measured by anemometer (Murthy Padmanabha 2009). Pollutant concentrations are recorded using High Volume Sampler by Improved West and Gaeke Method for SO_2, Jacob & Hochheiser Method for NO_2 and by Gravimetric Method for Particulate Matter (NAAQS monitoring and analysis guidelines 2011).

Air Quality Index is the method of reporting daily pollution levels to the general public with a view to integrate criteria air pollutants in the form of a number. As per United States Environmental Protection Agency, 1999 guidelines, AQI can be calculated with a two-step approach. First step is formation of sub-indices of each pollutant and the second consists of aggregation (breakpoint) of sub-indices.

As per Indian National Ambient Air Quality Standards (NAAQS) and epidemiological studies indicating risk of adverse health effects of specific pollutant, break point concentrations as

applicable to Indian conditions are used while calculating AQI.
In India, in order to reflect the status of air quality and its effects on human health , the range of index values has been designated as "Good (0-100)", "Moderate (101-200)", "Poor(201-300)", "Very Poor (301-400)" and "Severe (401-500)" (Khare and Nagendra 2007) (Kumar and Goyal 2011)

AQI is calculated for all three stations using following formula given by US EPA (Environmental Protection Agency) for four criteria air pollutants namely SO_2, NO_2, RSPM and SPM.

$$Ip = \left[(IHi - Ilo)/(BPHi - BPLo)\right](Cp - BpLo) + ILo \quad (1)$$

Where,

Ip = Index for pollutant *P*
Cp = Actual ambient concentration of pollutant p in μg/cum
BPHi = Breakpoint (as per Table 1) which is greater than or equal to *Cp*
BPLo = Breakpoint (as per Table 1) which is less than or equal to *Cp*
IHi = sub index value corresponding to *BPHi*
Ilo = sub index value corresponding to *BPLo*

Above formula gives Index value for each of the criteria pollutants and highest among them is considered as the overall AQI and the pollutant corresponding to highest index is declared as the responsible pollutant.

Table 1 Proposed sub index and breakpoint pollutant concentrations for Indian AQI

Sr. No.	Index Value	Descriptor	SO_2 (24 hr avg)	NO_2 (24 hr avg)	RSPM (24 hr avg)	SPM (24 hr avg)
1	0-100	Good [a]	0-80	0-80	0-100	0-200
2	101-200	Moderate [b]	81-367	81-180	101-150	201-260
3	201-300	Poor [c]	368-786	181-564	151-350	261-400
4	301-400	Very Poor [d]	787-1572	565-1272	351-420	401-800
5	401-500	Severe [e]	>1572	>1272	>420	>800

4. Soft computing tools

4.1 Artificial Neural Networks (*ANN*)

ANN is a computational system which is flexible and can be used as a potential tool to evaluate a nonlinear time series such as air quality. ANN's are required to be trained with some learning algorithm. It facilitates memorization of certain functions, classification of patterns and taking appropriate decisions. ANN has an ability of learning and generalization. Uniqueness of ANN lies

in its architecture, training algorithms and transfer / activation functions.

In this study a three layered feed forward back propagation network with a logsig and purelin as a transfer function is used. As shown in Fig. 3 the designed model structure consists of input, hidden and the output layer. Details of neural networks can be found in Bose and Liang (2000). For the present work, Neural Network toolbox provided by MATLAB is used.

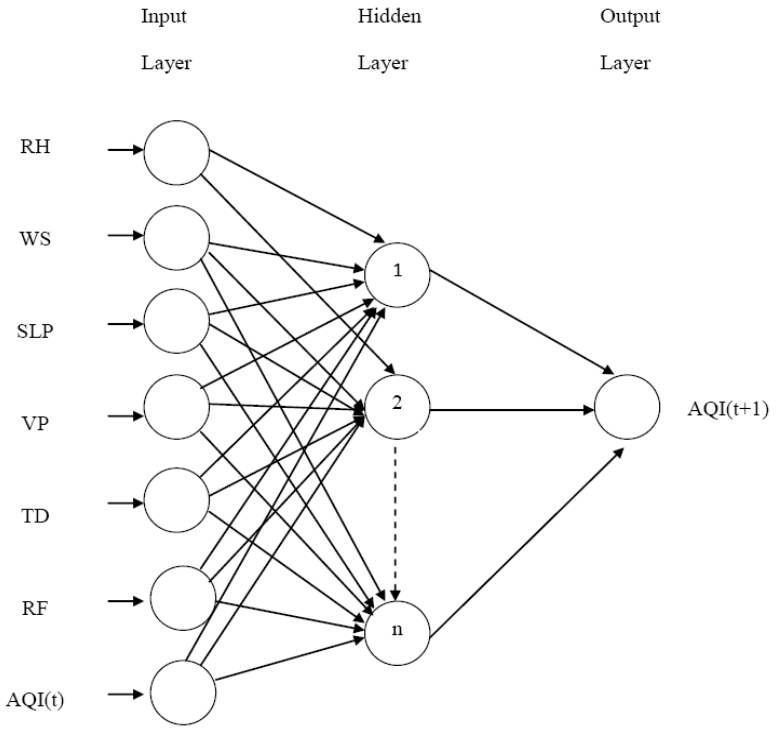

Fig. 3 The ANN model

4.2 Genetic Programming (*GP*)

Genetic Programming is an artificial intelligence technique which uses the principle of "Survival of the fittest" borrowed from the process of evolution occurring in nature. Its search strategy is based on Genetic Algorithms (GA) introduced by John Holland in 1960's. But unlike GA, GP's solution is a computer program or an equation as against a set of numbers in the GA; and hence GP can be conveniently used as a regression tool rather than optimization tool in GA. GP operates on parse trees rather than on bit strings as in a GA, to approximate the equation (in symbolic form) or computer program that best describes how the output relates to the input variables (Koza 1992).

In GP, a random population of individuals (equations or computer programs) is created, then the fitness of individuals is evaluated and the parents are selected out of these individuals. The parents are then made to yield the offsprings by following the process of reproduction, mutation

and crossover. The creation of offsprings is continued till a specified number of offsprings in a generation are produced and further till another specified number of generations are created. The resulting offsprings at the end of all this process (in the form of computer generated program or an equation) is the solution of the problem (Refer Fig. 4 for flowchart of GP).

In the present work, commercial software Discipulus is used to develop GP models. For all the GP models, control parameters such as initial population, crossover frequency & mutation frequency were updated in every run to evaluate the performance. Run was continued till there was no considerable reduction in fitness measure indicating generations without any further improvement. The best program with lowest error is considered as a final model.

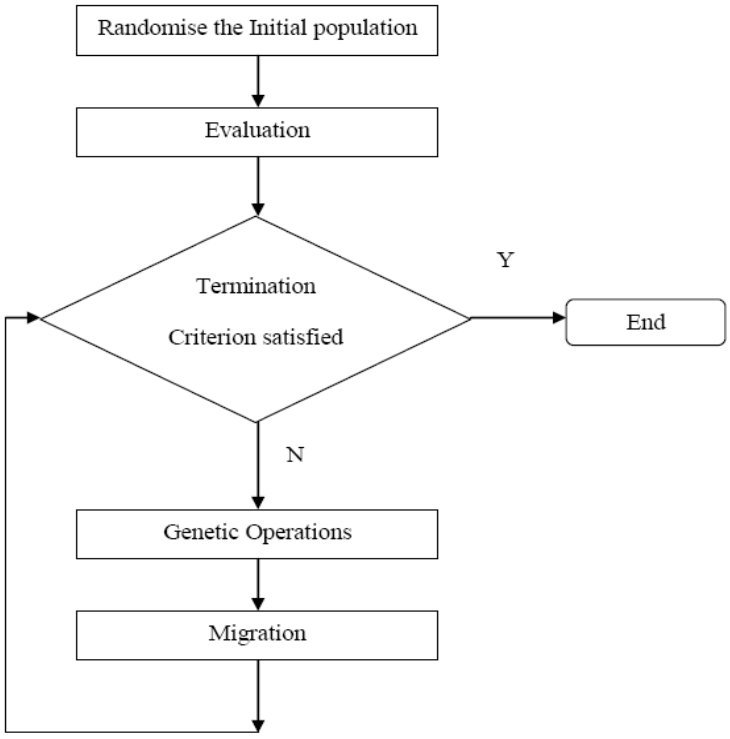

Fig. 4 Typical GP flowchart

5. Data preprocessing techniques

Data gathering methods are frequently loosely controlled mostly due to instrumental failure which may result into missing values, noise or impossible data combinations and the series is not evenly spaced in time. Misleading results can be produced with such kind of unprocessed data. Hence Data pre-processing is often an important step in data mining.

Data pre-processing include cleaning, normalization, transformation, feature extraction and

selection etc. There are number of data pre-processing techniques. They range from correlation analysis to some signal processing techniques such as spectral analysis, wavelet transform etc.

Data cleaning include cleaning the data by filtering in missing values, smoothing noisy data, identifying or removing outliers and resolving inconsistencies. Air quality data generally suffers from inconsistencies hence demand some cleaning method. Signal processing technique of spectral analysis can be applied to air quality data in order to identify the cyclic behavior and to remove the inconsistencies by differencing the series.

Present work discusses seasonal modeling. Periodic and cyclic observations of the meteorological parameters and pollutants are observed for four years which lead to periodic variation of AQI. This demands frequency based approach for analysis & treatment for error introduced due to time lag.

6. Auto correlation and spectral analysis

Autocorrelation refers to the correlation of time series with its own past and future values. It is also called as lagged correlation or serial correlation which refers to correlation between members of a series of numbers arranged in time remains in the same state from one observation to the next.

Autocorrelation complicates the application of statistical tests by reducing the number of independent observations. It can as well complicate the identification of significant covariance or correlation between time series.

Autocorrelation can be exploited for predictions. An auto correlated time series is predictable, probabilistically because future values depend on current and past values. There are three tools for assessing autocorrelation of time series namely the time series plot, the lagged scatter plot and the autocorrelation function.

The error can be analyzed and treated using frequency based technique of Spectral Analysis. Spectral analysis is concerned with the exploration of cyclical patterns of data. The purpose of the analysis is to decompose a complex time series with cyclical components into a few underlying sinusoidal (sine and cosine) functions of particular wavelengths. Then it will identify the correlation of sine and cosine functions of different frequency with the observed data. The sine and cosine functions are mutually independent (or orthogonal); thus may be the squared coefficients may be summed for each frequency to obtain the periodogram.

The periodogram values are computed as:

$$P_k = \text{sine coefficient}_k^2 + \text{cosine coefficient}_k^2 \times N/2 \tag{2}$$

where P_k is the periodogram value at frequency V_k and N is the overall length of the series. The periodogram values can be interpreted in terms of variance (sums of squares) of the data at the respective frequency or period. The periodogram values are plotted against the frequencies or periods.

AQI & meteorological parameters are considered as a multivariate time series and SA is proposed as a data pre-processing technique. Trial version XL stat is used to perform SA. Error due to time lag is calculated and the error frequency is estimated for all the seasonal models.

Seasonal models for all the three stations; Mumbai, Pune and Nasik are prepared using ANN and GP. Predictions are plotted on a time series. It was observed from the time series plot that there exists an autocorrelation. Positive time lag is also observed for almost all the predictions

which have introduced an error. This could be a result of autocorrelation. The error was analyzed and treated using Spectral Analysis.

7. Modeling strategy

Daily AQI is predicted on the basis of meteorological parameters and previous day's AQI, using ANN and GP on seasonal basis for the period of 2005-2008 for Mumbai, Pune and Nashik. Entire data is divided into four seasons namely Winter (January-February), Summer (March-May), Monsoon (June-September) and Post monsoon (October-December) (Murthy Padmanabha 2009). All these models are prepared with each tool namely ANN and GP. Thus there are total twenty four models. They involve seven inputs (six meteorological parameters and a previous value of AQI) and next day's AQI as output.

Thus all the models can be written as

$$AQI(t+1) = f(\mathrm{RH}(t), \mathrm{WS}(t), \mathrm{SLP}(t), \mathrm{VP}(t), \mathrm{TD}(t), \mathrm{RF}(t), \mathrm{AQI}(t))$$

Models prepared with ANN are named as ANN winter, ANN summer, ANN monsoon and ANN post monsoon and those prepared with GP are hereafter called as GP winter, GP summer, GP monsoon and GP post monsoon.

7.1 Criteria used for ANN based forecast models

For the present study, couple of trials for deciding data division were taken. Training and testing dataset, varying from 40%-85% (for training and remaining data for testing) were taken and found that 70% and 75% data for training and 30% and 25% of data respectively for testing yields better results. Hence the same data division was adopted. Network architecture comprises of inputs as number of input variables, one output neuron representing next day's AQI and hidden neuron as the smallest number of neurons which yield a minimum prediction error on the validation dataset (Khare and Nagendra 2007).

Each network had "logsig" and "pureline" as transfer function and mean squared error as a performance function. Training of the network is continued till a very low value of mean squared error was achieved in each case. Network weights are uniformly distributed in the range of -1 to 1.Training algorithm used in each case was Levernberg- Marquardt algorithm.

7.2 Criteria used for GP based forecast models

All the GP models were developed with the same data division as for ANN models so that the results can be compared. The control parameters of GP were updated in every run to judge the performance. Every run was continued till there was no further improvement. The best program with lowest fitness measure (error) is selected as the model. Average values of initial parameters selected for GP are Initial population size = 500, Mutation frequency = 95% and Crossover frequency = 50%. Mean squared error is considered to be the fitness criteria.

7.3 Performance evaluation

Testing performance was assessed by statistical parameters like correlation coefficient (*r*), root mean square error (RMSE), *d* statistics and mean bias error (MBE) between observed and the predicted values of AQI. Goodness of fit is measured using *r* and *d* and RMSE and MBE are the absolute error measures. Low values of errors (RMSE and MBE) and high value for *r* and *d* is treated as measure of satisfactory model performance.

8. Results and discussion

Present work predicts AQI on the basis of meteorological parameters and the previous values of AQI. Time series of meteorological variables are very noisy due to effects of climate variations. The whole time series of the dataset was used and it was ensured that there were no missing values and that all data were equally spaced in time (i.e., a regular time series). Regular time series are especially important in the case of leading indicators such as autocorrelation that estimate memory in time series. Interpolation can solve issues of missing values and irregular time series. But it can also result in spurious correlations. While checking interpolated records against the original time series it is required to ensure that the density of interpolated points is constant along the time series. Alternatively, points can also be dropped to obtain a regular time series. (e.g., some input parameters can be available on daily basis whereas others are available at an interval of 3 days then to make the series of all inputs uniform it is better to convert all data at an interval of 3 days). In addition, aggregation also may solve the issue of missing values, although at the cost of losing data. Here we did not need to aggregate our datasets because sampling in time scales that represented the characteristic time scale of the system; we simulated.

Thus regular time series are used to develop one day ahead seasonal AQI forecasting models using ANN and GP for all the three stations. The models were formulated as per the control parameters stated above .There were four seasonal models of each for Mumbai, Pune and Nashik using ANN and GP constituting twenty four models. The model results are compared in Table 2, 3 and 4 & time series plots are shown in Figs. 5, 6 and 7 for Mumbai, Pune and Nashik respectively.

8.1 Mumbai models

From the observation table it can be seen that, both the tools (ANN and GP) worked reasonably good for all the seasons and for all the stations. GP seems to capture the phenomenon even better than ANN as evident from the statistical evaluations. Decreased values of errors (RMSE and MBE) and increased values of *r* and *d* further support the statement. Negative MBE values except for GP Post monsoon (Mumbai) and ANN Post monsoon (Nahsik) indicate the tendency of the models to under predict. Whereas for Pune Models all MBE's are positive indicating the tendency of the models to overpredict. Out of the four Mumbai models maximum and minimum AQI recorded in winter was 252 and 25 respectively for testing data. Heavy fog during early mornings of winter supports the highest concentrations during winter. Rainfall during few days of winter results in the minimum pollutant concentrations and minimum values of AQI .GP could predict the maximum value as 245 as against 206 predicted by ANN. Thus GP could map the peaks very well but insufficient for predicting minimum value of AQI. ANN is closer to observed minimum values and it could predict minimum value as 66 as against 157 predicted by GP for Mumbai data.

Except for winter, GP could explain AQI approximately 71%, 79% and 72% accurately for summer, monsoon and post monsoon.

Table 2 Summary of ANN and GP models for AQI at Mumbai

Tools	ANN				GP				
Model	*r*	RMSE	*d*	MBE	Model	*r*	RMSE	*d*	MBE
ANN winter	0.456	33.026	0.571	-17.648	GP winter	0.541	26.019	0.641	-5.466
ANN summer	0.653	35.050	0.785	-7.683	GP summer	0.714	34.046	0.837	-3.441
ANN monsoon	0.737	35.924	0.813	-12.681	GP monsoon	0.791	31.052	0.876	-6.066
ANN post monsoon	0.686	36.666	0.814	-4.382	GP post monsoon	0.727	34.225	0.833	1.983

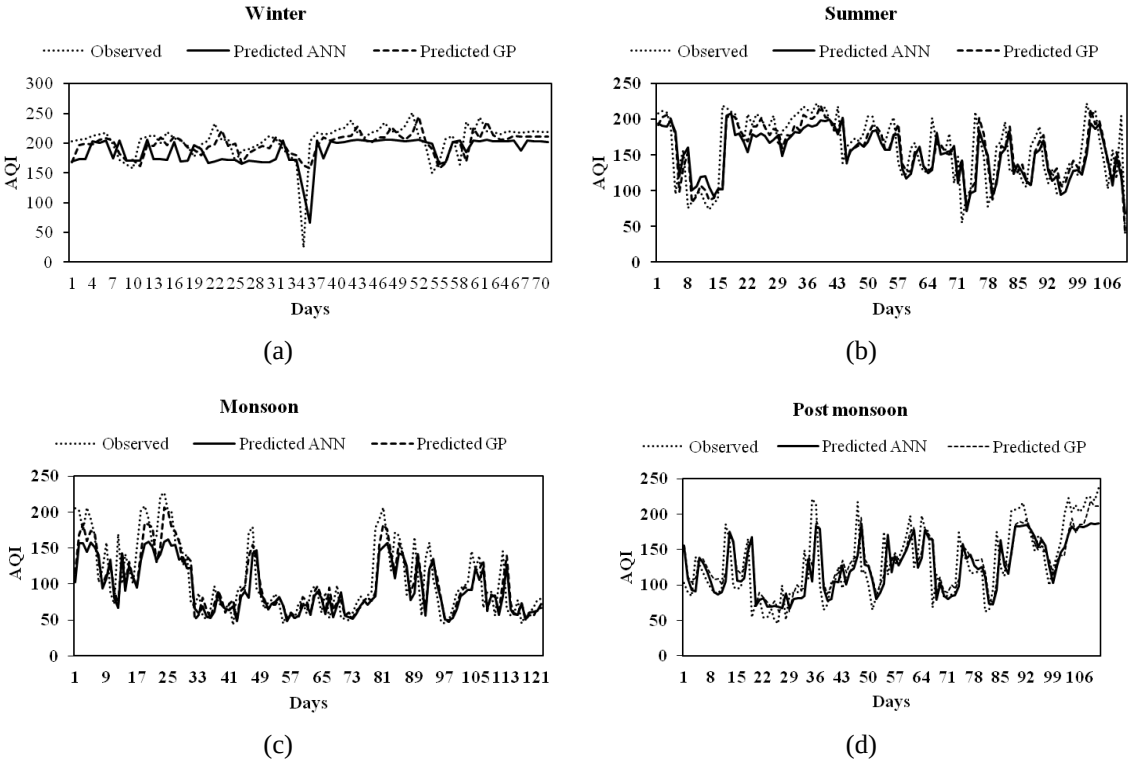

Fig. 5 Seasonal comparison of observed and predicted AQI for Mumbai

8.2 Pune models

For Pune it can be seen that GP works better than ANN which is evident from results. GP could also follow the peak values better than ANN, which is evident from monsoon and post monsoon models. Maximum value of AQI is recorded as 271 for summer and minimum value is recorded as 36 for monsoon and the responsible pollutant in both the cases is RSPM. The model predicted maximum values are 205 and 207 by ANN and GP models respectively. Whereas minimum values

predicted by ANN and GP are 83 and 51 respectively. In both the cases GP is closer to observed values than ANN.

From the statistical evaluations, it can be seen that GP could explain AQI by approximately 83%, 82%, 80% and 69% for monsoon, postmonsoon, summer and winter. The reason for highest value recorded in early summer is the transition time i.e., end of winter and beginning of summer. Due to lesser air movements the particles gets traped and increases the concentration. Whereas in monsoon, rainfall act as a cleansing mechanism and washes away all the pollutants hence minimum values are observed.

Table 3 Summary of ANN and GP models for AQI at Pune

Tools	ANN				GP				
Model	*r*	RMSE	*d*	MBE	Model	*r*	RMSE	*d*	MBE
ANN winter	0.486	37.531	0.804	7.891	GP winter	0.694	30.725	0.864	2.666
ANN summer	0.734	45.138	0.804	13.325	GP summer	0.799	38.004	0.876	1.458
ANN monsoon	0.41	28.07	0.64	1.6	GP monsoon	0.58	23.75	0.75	2.22
ANN post monsoon	0.732	35.788	0.833	2.444	GP post monsoon	0.819	30.346	0.884	1.468

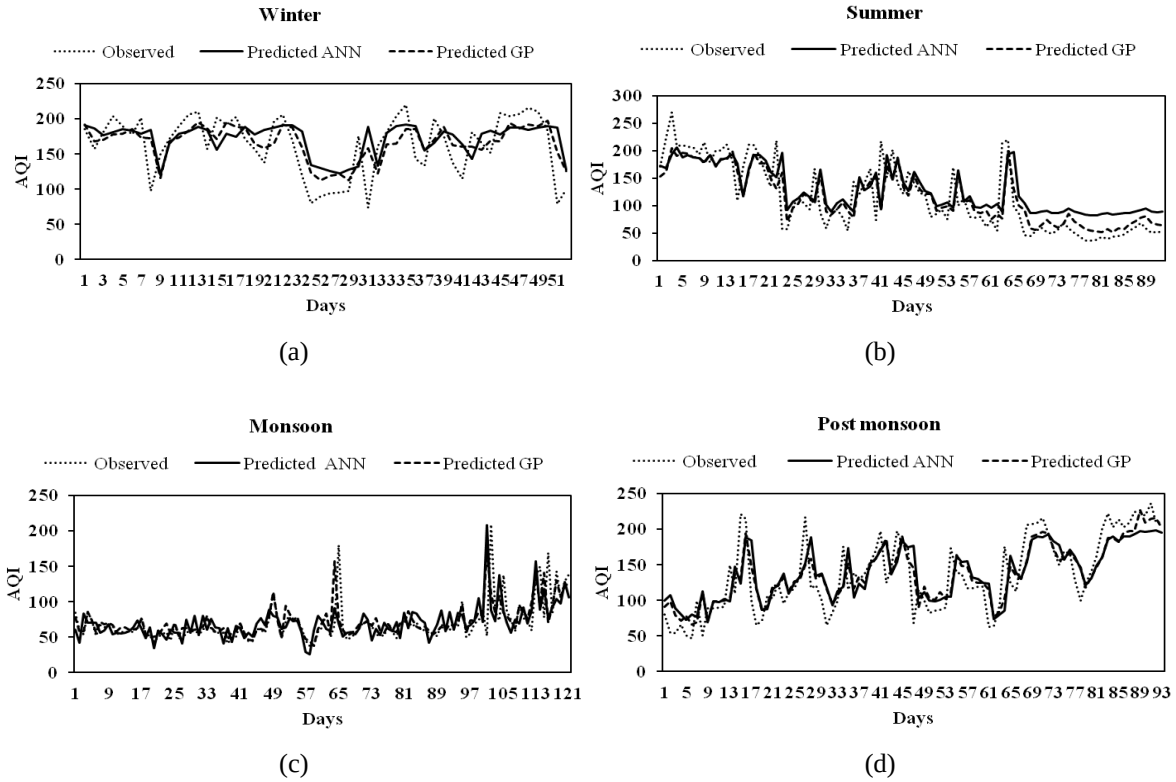

Fig. 6 Seasonal **c**omparison of observed and predicted AQI for pune

8.3 Nashik models

For Nashik data both the tools worked reasonably well but GP was a shade better than ANN which is evident from results.

GP models could predict the peak values better where ANN is found to be weak especially for monsoon model. Maximum observed AQI is 252 for winter model and minimum AQI is recorded as 32 for monsoon model. Maximum predicted AQI is 250 by GP and 247 by ANN whereas minimum predicted AQI is 28 by ANN and 37 by GP. Thus predictions by GP are much closer to observed values compared to ANN. Statistical performance of all the models indicate that GP could map the phenomenon approximately 84%, 72%, 70% and 51% for monsoon, post monsoon, summer and winter models.

Table 4 Summary of ANN and GP models for AQI at Nashik

Tools	ANN				GP				
Model	*r*	RMSE	*d*	MBE	Model	*r*	RMSE	*d*	MBE
ANN winter	0.506	26.325	0.623	-6.307	GP winter	0.507	25.800	0.898	-2.433
ANN summer	0.665	35.657	0.784	-12.507	GP summer	0.706	31.166	0.830	-3.375
ANN monsoon	0.785	36.082	0.825	-17.138	GP monsoon	0.839	28.026	0.907	-4.305
ANN post monsoon	0.662	37.417	0.809	4.550	GP post monsoon	0.719	33.271	0.841	-0.077

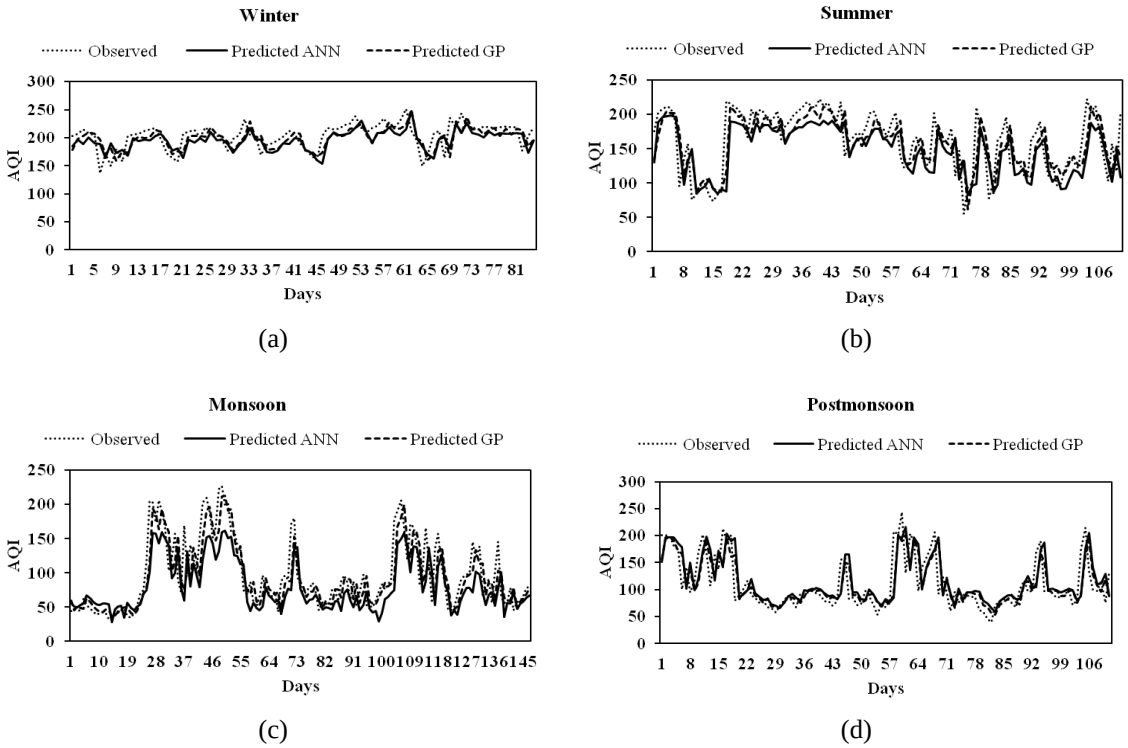

Fig. 7 Seasonal comparison of observed and predicted AQI for Nashik

9. Analysis and treatment of phase difference in prediction process

From the above seasonal time series plots, it can be seen that there is a prediction lag between observed and predicted variables for all the models prepared with both the tools.

Conventional methods of rescaling or standardization of the training data are insufficient hence pre-processing methods from the perspective of signal analysis are also crucial as the time series of meteorological parameters could be viewed as quasi periodic signal which is contaminated by noises. Lag in prediction here in this case is due to autocorrelation.

Spectral analysis is therefore executed which consists of plotting a periodogram. Periodogram analysis for air pollution data has advantage of quantification and characterization of the periodic behavior of air quality time series (Dilmaghani 2007). Air pollution data contains periodic patterns. Diurnal fluctuations of meteorological conditions, seasonal fluctuations in solar radiation, earth's daily rotation and human activities may cause early seasonal, daily periodicities in the time series. These periodicities can be analyzed by the study of periodogram. Periodogram not only drastically simplified in finding periodicities in time series but also helped to extract information from time series.

In signal processing Fourier transform is decomposing a signal into its frequencies and amplitude while mapping the time series in the frequency domain. Frequency domain graph carries important information about a signal (time series), amplitude and phase. This information helped to regenerate the original signal (data) from frequency spectrum. It is very important concept in air quality data because it helps to refill the gaps and missing values after redeveloping the data .Moreover analyzing the data in the frequency domain is extremely useful because it not only reveals periodicities in input data but also the relative strength of any periodic component. Spectral Analysis (SA) is a technique of time series analysis. It allows transforming a time series into its coordinates in the space of frequencies and then analyzes its characteristics.

For the present work, when the time lag is observed an original time series of input variables is analyzed. It was found non stationary. Non stationary time series are complex hence are made stationary by differencing and converting each *i*′th element of the series into its differenced from the (*i*-*k*)″th element. Thus inconsistencies are removed from the data. Trend of error is estimated by signal extraction technique of spectral analysis.

Table 5 indicates the average repetition time of error cycle. Average error cycle for winter, summer, monsoon and post monsoon is 2, 2.24, 4.44 and 3.67 days respectively. The error cycle is analyzed and found to be repeated after some specific time interval for different seasons. The error cycle repetition time suggests the time interval for which the readings are affected due to time lag. In winter the error is more frequent as compared to other seasons which suggest that the winter predictions are affected by time lag and auto correlation compared to predictions in other seasons. This is also evident from the model results. Performance of all the models except winter before time lag correction is better. The reason for higher effect of error could be the characteristics of the season, accordingly changes in the meteorology and its effect on pollutant concentrations. Winter has worst air quality as the dispersion is limited. The situation worsens as the winter progresses towards summer because of the photochemical reactions progressing in summer. The situation is better in monsoon and post monsoon as rainfall is a cleansing mechanism. The changes in wind velocity and reversal of its direction carry the pollutants away from sources as well as increase the possibilities of dilution of concentration of pollutants also. Hence the correction is applied and then with the new series all the models are evaluated using soft computing tools.

The exception to this is observed for Mumbai. Highest AQI for Mumbai is observed for

monsoon season. The possible reason for highest AQI for Mumbai could be below average rainfall for the years 2006, 2007 and 2008 as observed from recoded data. (The average rainfall in Mumbai is 242 cm and the rainfall recorded is around 100 cm in these years).The trend of AQI for Mumbai in Monsoon season needs to be studied further in depths for the years in which average rainfall has occurred.

The correction is applied and with the new series all the models are evaluated using soft computing tools. The new models prepared using ANN are named as ANN Winter1, ANN Summer1, ANN Monsoon1, ANN Post monsoon1 and those prepared with GP are called as GP Winter 1, GP Summer 1, GP Monsoon 1, GP Post monsoon 1.

It was found that when the data preprocessing technique are coupled with soft computing tools, the relationship between inputs and outputs is better mapped. The time lag is removed and the performance of all the models is significantly improved which is evident from the results.

Table 5 Error cycle estimation using SA

Sr. No.	City	Model	Average Error cycle repetition time for soft computing tools (ANN & GP)
1	Mumbai	Winter	2.088
		Summer	2.115
		Monsoon	3.813
		Post monsoon	6.167
2	Pune	Winter	2.080
		Summer	2.486
		Monsoon	4
		Post monsoon	2.325
3	Nashik	Winter	2
		Summer	2.135
		Monsoon	5.5
		Post monsoon	2.523

9.1 Mumbai models after time lag correction

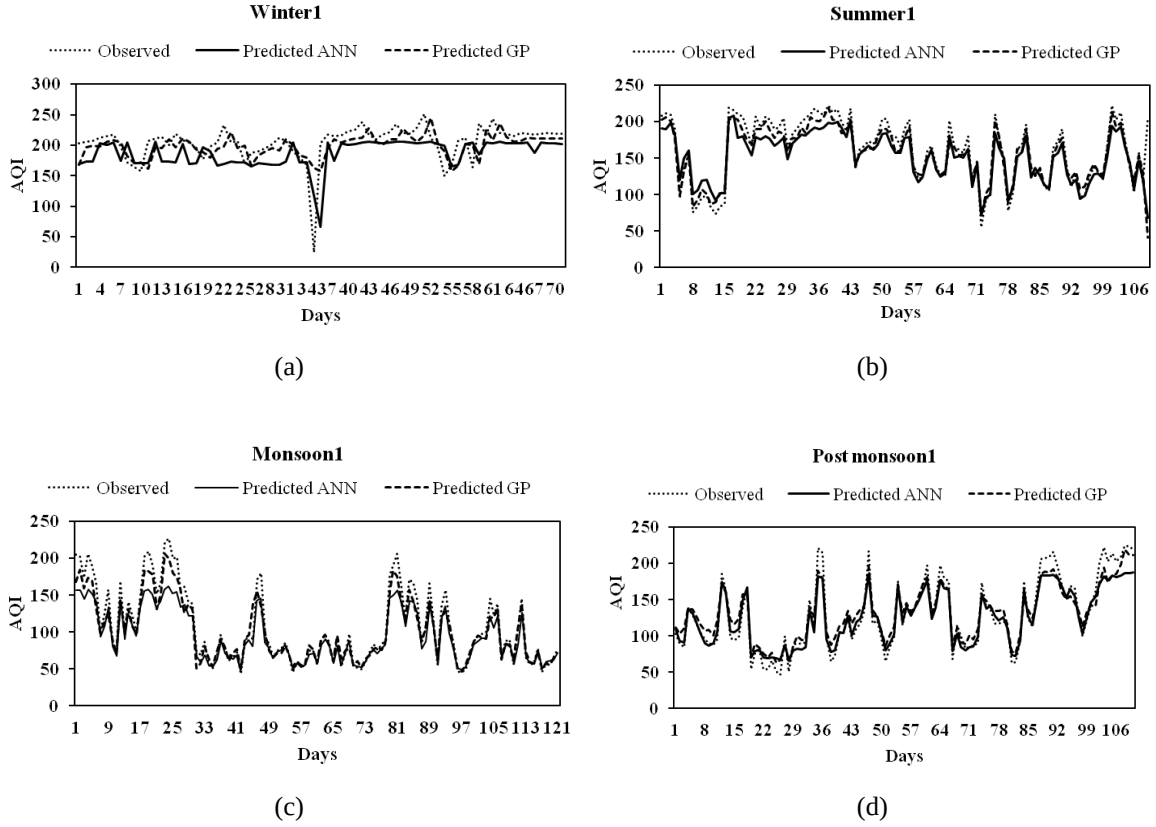

Fig. 8 Seasonal comparison of observed and predicted AQI for Mumbai after time lag correction

Table 6 Summary of ANN and GP models for AQI at Mumbai after time lag correction

Tools	ANN				GP				
Model	*r*	RMSE	*d*	MBE	Model	*r*	RMSE	*d*	MBE
ANN winter1	0.790	35.825	0.643	-20.030	GP winter1	0.850	31.430	0.648	-7.831
ANN summer1	0.910	20.257	0.934	-9.167	GP summer1	0.921	18.840	0.940	-4.870
ANN monsoon1	0.960	20.162	0.955	-12.86	GP monsoon1	0.970	14.460	0.970	-6.060
ANN post monsoon1	0.960	14.417	0.979	-3.661	GP post monsoon1	0.971	15.741	0.970	2.870

9.2 Pune models after time lag correction

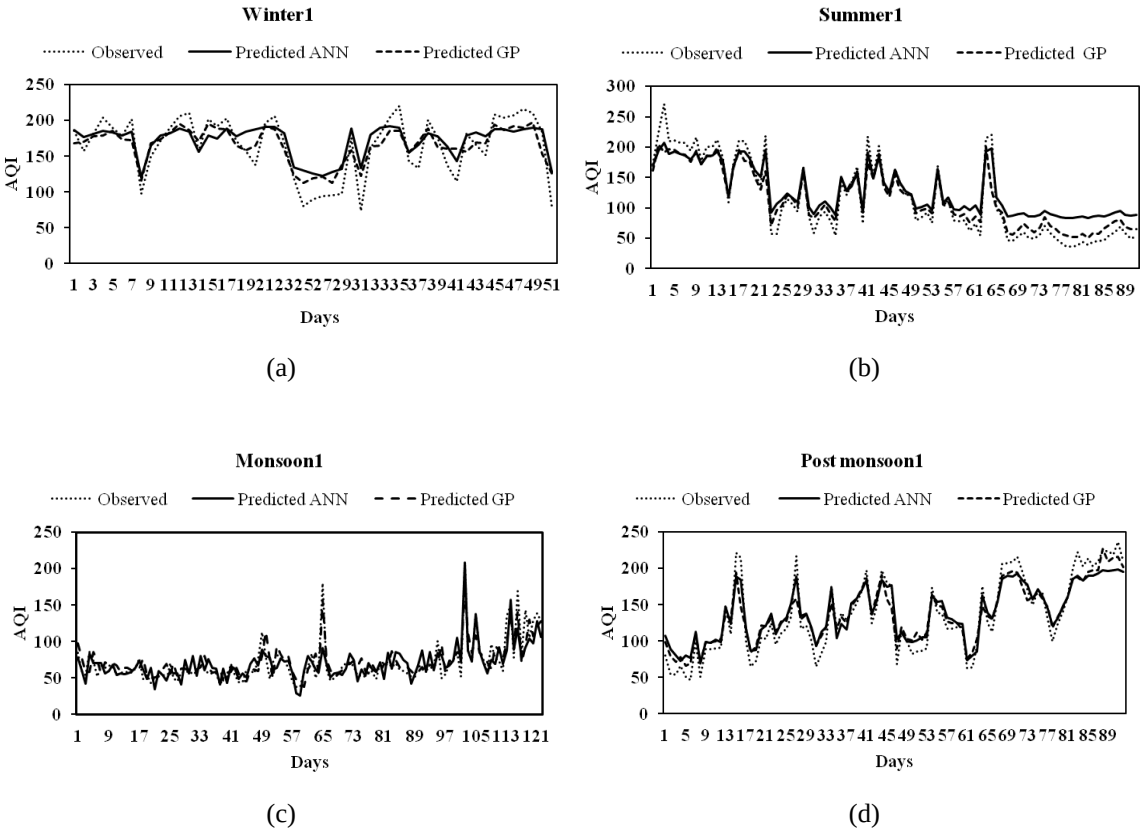

Fig. 9 Seasonal comparison of Observed and Predicted AQI for Pune after time lag correction

Table 7 Summary of ANN and GP models for AQI at Pune after time lag correction

Tools	ANN				GP				
Model	*r*	RMSE	*d*	MBE	Model	*r*	RMSE	*d*	MBE
ANN winter1	0. 901	23.630	0.933	6.140	GP winter1	0.910	21.610	0.940	0.891
ANN summer1	0.970	26.561	0.934	11.441	GP summer1	0.991	20.662	0.961	-0.201
ANN monsoon1	0.825	16.105	0.895	2.244	GP monsoon1	0.829	16.128	0.896	1.971
ANN post monsoon1	0.980	27.990	0.919	1.211	GP post monsoon1	0.961	28.331	0.914	2.661

9.3 Nashik Models after time lag correction

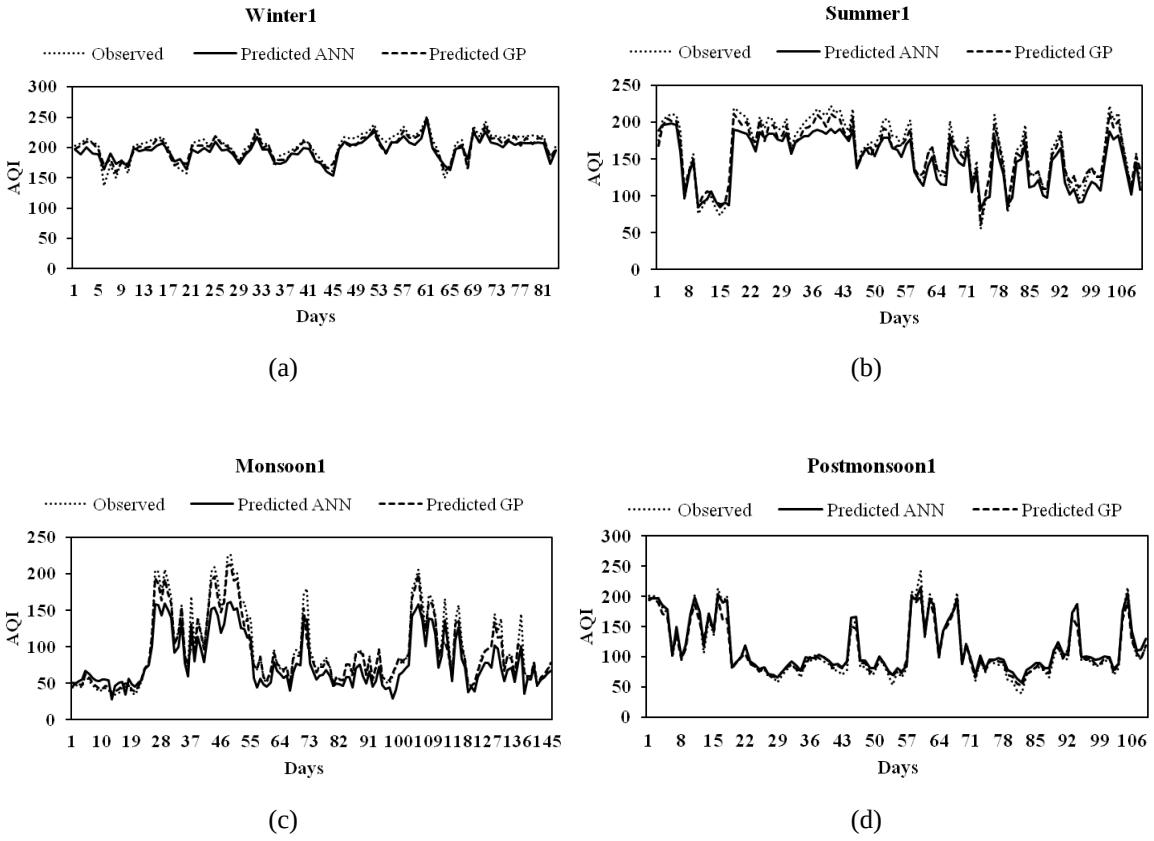

Fig. 10 Seasonal comparison of observed and predicted AQI for Nashik after time lag correction

Table 8 Summary of ANN and GP models for AQI at Nashik after time lag correction

Tools	ANN				GP				
Model	*r*	RMSE	*d*	MBE	Model	*r*	RMSE	*d*	MBE
ANN winter1	0.730	26.781	0.791	-10.110	GP winter1	0.750	25.512	0.813	-6.161
ANN summer1	0.980	25.824	0.892	-13.561	GP summer1	0.981	22.050	0.930	-4.470
ANN monsoon1	0.916	25.251	0.923	-17.521	GP monsoon1	0.840	28.023	0.912	-4.032
ANN post monsoon1	0.992	16.698	0.979	2.591	GP post monsoon1	0.981	17.714	0.960	-2.040

Thus AQI prediction models can be prepared with soft computing tools coupled with data preprocessing techniques which would tell us about the air quality on the next day.

10. Conclusions

In order to get the realistic seasonal picture about the air quality, seasonal modeling approach is adopted. ANN and GP models are developed to predict one day ahead air quality index for three major cities of Maharashtra state of India. Time lag was observed for all twenty four models. Time lag introduced due to high autocorrelation was characterized by using Spectral analysis. New time series is generated and models are reevaluated. All the models performed well in testing. Combined models of Spectral analysis and GP was found to be superior to Spectral analysis and ANN models. Hence SA coupled with GP can be used as a tool in case of air pollution episode predictions.

References

Bose, N.K. and Liang, P. (2000), *Neural Network Fundamentals with Graphs, Algorithms and Applications*, Tata McGraw-Hill Publication, Delhi, India.

Brunelli, U., Piazza, U. and Pignato, L. (2007), "Two-day ahead prediction of daily maximum concentrations of SO_2, O_3, PM_{10}, NO_2, CO in the urban area of Palermo, Italy", *Atmosp. Environ.*, **41**(14), 2967-2995.

Chen, J.-L., Islam, S. and Biswas, P. (1998), "Nonlinear dynamics of hourly ozone Concentrations: nonparametric short term prediction", *Atmosp. Environ.*, **32**(11), 1839-1848.

Comrie, A.C. and Diem, J.E. (1999), "Climatology and forecast modeling of ambient carbon monoxide in Phoenix, Arizona", *Atmosp. Environ.*, **33**(30), 5023-5036.

Davis, J.M. and Speckman, P. (1999), "A model for predicting maximum and 8 h average ozone in Houston", *Atmosp. Environ.*, **33**(16), 2487-2500.

Dilmaghani, S. (2007), "Spectral analysis of air quality data", Dissertation Report; University of Southern California, CA, USA.

Draxler, R.R. (2000), "Meteorological factors of ozone predictability at Houston, Texas", *J. Air Waste Manag. Assoc.*, **50**(2), 259-271.

EPA, (1999), *Air Quality Index Reporting*; *Final Rule*, Federal Register, Part III, 40 CFR Part 58.

Gardner, M.W. and Dorling, S.R. (2000), "Artificial neural networks: The multilayer perceptron: A review of applications in the atmospheric sciences", *Atmosp. Environ.*, **32**(14-15), 2627-2636.

Grivas, G. and Chaloulakou, A. (2005), "Artificial neural network models for prediction of PM_{10} hourly concentrations, in the Greater Area of Athens, Greece", *Atmosp. Environ.*, **40**(7), 216-1229.

Hajek, P. and Olej, V. (2009), "Air quality indices and their modelling by hierarchical fuzzy inference systems", *WSEAS Transactions on Environment and Development*, **10**(5), 661-672.

Hubbard, M. and Cobourn, W.G. (1998), "Development of a regression model to forecast ground-level ozone concentration in Louisville, KY", *Atmosp. Environ.*, **32**(14-15), 2637–2647.

Jiang, D., Zhang, Y., Hu, X., Zeng, Y. and Tan, J. and Shao, D. (2004), "Progress in developing an ANN model for air pollution index forecast", *Atmosp. Environ.*, **38**(40), 7055-7064.

Kassomenos, P.A., Kelessis, A., Petrakakis, M., Zoumakis, N., Christidis Th. and Paschalidou, A.K. (2012), "Air quality assessment in a heavily polluted urban Mediterranean environment through air quality indices", *Ecol. Indicators*, **18**, 259-268.

Khare, M. and Nagendra, S.A. (2007), "Artificial neural networks in vehicular pollution modelling", *J. Stud. Comput. Intell.*, 41-45.

Koza, J.R. (1992), *Genetic Programming on the Programming of Computers by Means of Natural Selection*,

A Bradford Book, MIT Press.
Kumar, A. and Goyal, P. (2011), "Forecasting of air quality in Delhi using Principal component regression technique", *Atmosp. Pollut. Res.*, **2**, 436-444.
Kyrkilis, G., Chaloulakou, A. and Kassomenos, P.A. (2007), "Development of an aggregate Air Quality Index for an urban Mediterranean agglomeration: Relation to potential health effects", *Environ. Int.*, **33**(5), 670-676.
Mayer, H., Jutta, H., Dirk, S. and Dieter, A. (2008), "Evolution of the air pollution in SW Germany evaluated by the long-term air quality index LAQx", *Atmosp. Environ.*, **42**(20), 5071-5078.
Milanchus, M.L., Rao, T. and Zurbenko, I.G. (1998), "Evaluating the effectiveness of ozone Management efforts in the presence of meteorological variability", *J. Air Waste Manag. Assoc.*, **48**(3), 201-215.
Mohan, M. and Anurag, K. (2007), "An Analysis of Annual and Seasonal Trends of Air Quality Index of Delhi", *Environ. Monitor. Assess.*, **131**(1-3), 267-277.
Murthy Padmanabha, B. (2009), *Environmental Meteorology*, I.K. International Publishing House Pvt. Ltd., New Delhi, India.
Perez, P. and Reyes, J. (2006), "An integrated neural network model for PM10 forecasting", *Atmosp. Environ.*, **40**(16), 2845-2851.
Pires, J.C.M., Alvim- Ferraz, M.C.M., Pariera, M.C. and Martins, F.G. (2011), "Prediction of troposphere ozone concentration: Application of a methodology based on Darwin's Theory of Evolution", *Expert Syst. Appl.*, **38**(3), 1903-1908.
Pun, B.K., Louis, J.F., Pai, P., Seigneur, C., Altshuler, S. and Franco, G. (2000), "Ozone formation in California's San Joaquin Valley: A critical assessment of modeling and data needs", *J. Air Waste Manag. Assoc.*, **50**(6), 961-971.
Raga, G.B. and Le Moyne, L. (1999), "On the nature of air pollution dynamics in Mexico City – I. Nonlinear analysis", *Atmosp. Environ.*, **30**(23), 3987-3993.
Rao, C.V.C., Chelani, A.B., Phadke, K.M. and Hasan, M.Z. (2002), "Formation of an air quality index in India", *Int. J. Environ. Stud.*, **59**(3), 331-342.
Roth, P.M. (1999), "A qualitative approach to evaluating the anticipated reliability of a photochemical air quality simulation model for a selected application", *J. Air Waste Manag. Assoc.*, **49**(9), 1050-1059.
Russell, A. and Dennis, R. (2000), "NARSTO critical review of photochemical models and Modeling", *Atmosp. Environ.*, **34**(12-14), 2283-2324.
Salcedo, R.L.R., Alvim, M.C.M., Alves, C.A. and Martins, F.G. (1999), "Time-series analysis of air pollution data", *Atmosp. Environ.*, **33**(15), 2361-2372.
Sebald, L., Treffeisen, R., Reimery, E. and Hies, T. (2000), "Spectral analysis of air pollutants. Part 2: Ozone time series", *Atmosp. Environ.*, **34**(21), 3503-3509.
Sharma, M., Pandey, R., Maheshwari, M., Sengupta, B., Shukla, B.P., Gupta, N.K. and Johri S. (2003), "Interpretation of air quality data using an air quality index for the city of Kanpur, India", *J. Environ. Eng. Sci.*, **2**(6), 453-462.
Thomas, S. and Jacko, R.B. (2007), "Model for forecasting expressway fine particulate matter and carbon monoxide concentration: Application of regression and neural network model", *J. Air Waste Manag. Assoc.*, **57**(4), 480-488.
Thompson, M.L., Reynolds, J., Cox, L.H.M, Guttorp, P. and Sampson, P.D. (2001), "A review of statistical methods for the meteorological adjustment of tropospheric ozone", *Atmosp. Environ.*, **35**(3), 617-630.
Tikhe Shruti, S., Khare, K.C. and Londhe, S.N. (2013), "Forecasting criteria air pollutants using data driven approaches: An Indian case study", *IOSR J. Environ. Sci., Toxicol. Food Technol.*, **3**(5), 01-08.
URL: http://en.wikipedia.org/wiki/Mumbai (Accessed on February 2014)
URL: http://safar.tropmet.res.in (Accessed on February 2014)
URL: http://www.dnaindia.com (Accessed on February 2014)
URL: http://www.hoparoundindia.com (Accessed on February 2014)
URL: http://www.imdpune.gov.in (Accessed on February 2014)
URL: http://www.mpcb.gov.in (Accessed on February 2014)
URL: http://www.xlstat.com (Accessed on February 2014)

Van den Elshout, S., Leger, K. and Nussio, F. (2008), "Comparing urban air quality in Europe in real time - a review of existing air quality indices and the proposal of a common alternative", *Environ. Int.*, **34**(5), 720-726.

Wong, T.W., Tam, W.W.S., Yu, I.T.S., Lau, A.K.H., Pang, S.W. and Wong, A.H.S. (2013), "Developing a risk based air quality health index original research article", *Atmosp. Environ.*, **76**, 52-58.

Zadeh, L. (1994), *Fuzzy Logic, Neural Networks and Soft Computing*, Communications of the ACM, **37**(3), 77-84.

WL