Shruti Tikhe, Kanchan Khare, S. N. Londhe

6 Air quality forecasting using soft computing techniques

Abstract: Air pollution is a fundamental problem among the critical challenges of modern society. Acute as well as long-term exposure may pose serious health problems. Real-time air quality predictions are necessary to forecast the pollutant concentrations. This information can be used to issue early air quality warnings that allow government and people to take preventive measures.

The authors offer application of soft computing techniques of artificial neural networks (ANN) and genetic programming (GP) for air quality parameter forecasting, a few time steps in advance. Seasonal forecasting models are also useful for proactive measures. Availability of continuous field data sometimes is a challenge. It is time-consuming and expensive. Modeling options are suggested for situations when there are constraints about data availability and the urgent need to understand and broadcast the air quality.

A case study of Pune, a metropolitan city in India, demonstrating the strengths of GP is presented. We have evaluated the models by a correlation coefficient, *d*-statistics, root mean squared error, and mean bias error.

Keywords: Air quality, soft computing, ANN, GP

6.1 Introduction

Air is a fundamental need for survival and development of all lives on the Earth. It has major impact on health and also influences the economic development of the country [18]. Air quality describes the chemical state of the atmosphere at a particular time and place. Rapid industrialization, development of cities, and concerned human activities are the major causes for an alarming increase in air pollution levels and degradation of the air quality especially in almost all the metro cities of the world [23]. Short-term exposure to high levels of pollution may pose severe health impacts such as eye irritation, difficulty in breathing, and cardiovascular and pulmonary health effects. Health concerns associated with long-term exposure can be

---

Shruti Tikhe, DTK Hydronet Solutions, Bavdhan, Pune, Maharashtra State, India

Kanchan Khare, Civil Engineering Department, Symbiosis Institute of Technology Lavale, Pune, Maharashtra State, India

S. N. Londhe, Department of Civil Engineering, Vishwakarma Institute of Information Technology Kondhwa, Pune, Maharashtra State, India

listed as the threat of cancer, risk of premature death, and also damage to the body's immune, respiratory, neurological, and reproductive systems. Senior citizens, kids, and individuals with preexisting heart and lung maladies along with diabetics are at greater risk for air pollution-related health effects [19]. The World Health Organization has issued guidelines related to permissible levels of pollutants with an objective to protect human health and the environment. In spite of significant progress in understanding the process of emission and fates of the air pollutants as well as in decreasing their ambient levels in urban localities for the last five decades, air pollution is estimated to destroy the lives of about 4 million people every year worldwide [34]. It is therefore worthwhile to model and forecast the air quality.

6.2 Issues and challenges

Air quality is described by the concentrations of major air pollutants such as sulfur dioxide, nitrogen oxides, and respirable suspended particulate matter (SO_2 , NO_x , and RSPM) that reflect the status of air pollution in the environment [36]. Air quality assessment studies are essential to determine the trends and effects of air pollutants on human lives, animals, soil, and environment. Source contribution to pollution can also be identified and accordingly controlling strategies can be decided [11].

Air pollution comprises a mixture of gases and particles in harmful amounts that are released into the atmosphere due to either natural or human activities. The leading sources of air pollution are natural as well as anthropogenic. Natural phenomena such as volcanic eruptions and forest fires release harmful substances such as SO_2 , CO_2 , NO_2 , and CO .

Anthropogenic sources include the pollutants entering from a fixed point, mobile sources, and area sources in addition to the burning of fossil fuel. Pollutants emitted from human-made sources include hydrogen, oxygen, nitrogen, sulfur, metal compounds, and particulate matter.

They affect badly on human health and the environment. The global impact of air pollution is on climate change and particularly on global warming.

During the last five decades, scientific research has been carried out to model as well as forecast the air quality. Various methods and techniques have been evolved for modeling of air pollutants. For effective air quality management, monitoring and record-keeping of the pollutant concentration data become a vital prerequisite. Many countries have their real-time air quality monitoring as well as forecasting systems. Developed countries have their established air quality monitoring system, but the issue is with the developing countries where availability of the continuous data is a significant challenge.

6.3 Air quality modeling

It has been observed that air quality assessment and modeling has been carried out using a variety of approaches. The complexity of the problem and the availability of the resources decide the choice of method [6]. Traditionally, hard computing approaches have been used to carry out air quality modeling and forecasting. They include either physical, deterministic, stochastic, statistical, or numerical models [30]. Deterministic models require emission inventory and meteorological variables to calculate the pollutant concentrations. These models have been used traditionally to calculate air pollutants in urban areas [20].

Physical models are costly, data-intensive, and challenging to construct and operate; as a result, their use has been restricted only to understand the complex urban dispersion phenomenon [1]. Majority of the dispersion models developed to evaluate and predict the pollutants in urban areas are “causal” in nature and therefore fail to predict the “extreme” concentrations [9]. In air quality management, forecasting and broadcasting of the peak concentrations (extremes) are very important to give sufficient warning to the citizens. Statistical techniques represent only a specific monitoring station; as a result, they cannot be extended to other regions with different meteorological conditions [14]. The interrelationship between human, climate, and air pollution is extremely complex; as a result, it is very difficult to be represented in deterministic models without developing a separate statistical model [6]. As meteorology affects air quality and this relationship is complex and nonlinear, data-driven approaches are suitable; however, they are also site-specific. They provide tools to facilitate the conversion of the data sets into a convenient form and convey a better understanding of the processes hidden in these data sets by expressing the mathematical statement of the relations between various parameters in the data [2]. Soft computing tools have been developed as an alternative to deterministic and physical models.

6.4 Forecasting of air quality parameters

Air quality parameters are stochastic in nature; thus, short-term predictions are possible based on past meteorological data [22]. Soft computing tools are modern forecasting approaches. Resource availability and accuracy requirements decide the choice of the tool. Chosen tool should be robust enough to take into account the data fluctuations without compromising the accuracy. Extensive experimentation is required to identify such a tool [12].

6.4.1 Artificial neural networks

There are many data-driven approaches known in computer science today. One of the most popular approach is artificial neural networks (ANNs). ANNs are biologically inspired, parallel computational tools. They consist of simple highly interconnected processing elements that process the input similar to the human brain [28]. ANN is broadly classified as a feed-forward neural network and a recurrent neural network. Figure 6.1 shows a feed-forward network [15], which typically has three layers, namely input, output and hidden layer. Input layer neurons receive the input variables from the problem, and output layer neuron/neurons yield the network output (desired result). ANN works similar to the human brain on two fronts. First one is acquiring knowledge from the input data via the network of neurons through a learning process and interneuron connection strengths are utilized to store the acquired knowledge. The weights [between the input layer and hidden layer as well as hidden layer and output layer] represent the relative strength of specific input parameters in the computation of output. An artificial neuron weighs and sums up the inputs and compares the results with a predefined threshold. Iterations are carried out until the stopping criteria are reached. The network-generated output heavily depends on the type of activation functions (also called transfer or squashing function) used. Three types of transfer functions are generally used, namely, step, ramp, and sigmoid function. The sigmoid function is the most

Figure 6.1: The feedforward network.

Source: The ASCE Task Committee, 2000 [15].

commonly used transfer function for air quality modeling [13]. The recurrent network has a feedback path which makes it sequential rather than combinatorial. Once cyclic connections are included, the network becomes a nonlinear dynamic system. Backpropagation is one of the most widely used algorithms in the feed-forward type of networks, which is based on the steepest descent method.

The performance of ANN models entirely depends on the quality and the quantity of the data. ANNs were considered as black box earlier but some researchers have shown that the network incorporates all of the physics underlying the process since they are embedded in the data and ANN is not merely playing with numbers [25]. Similarly, the works of Londhe and Shah [10] and Londhe and Panchang [26] throw a light on the work of knowledge extraction from the trained network. Deciding ANN architecture is a complex process that demands the skill of the modeler. There is no fixed rule to determine the data division and the number of neurons in the hidden layer. It has been reported that ANN models may suffer from overfitting due to the addition of too many hidden neurons or executing a large number of iterations. Some researchers have found that ANNs have a capability of prediction of extreme events when combined with other techniques such as wavelets [4].

6.4.2 Genetic programming

Genetic programming (GP) is an evolutionary data-driven modeling approach that is similar to genetic algorithms (GA) and follows the Darwinian principle of the survival of the fittest and obtains the solution to the problem through the process of crossover, mutation, and reproduction. GP can be conveniently used as a regression tool as the solution of GP is a computer program or an equation as against a set of numbers in GA [24].

The process of GP starts with a population of randomly generated computer program initially constructed from the data set on which genetic operations are performed using function and terminal set. Function set comprises the operator to be used such as addition, subtraction, logarithm, and square root. Terminal set consists of values such as inputs, constants, and temporary variables on which the function set operates. From the population so generated, four programs are randomly selected, and a tournament is conducted.

GP measures the performance of each program, and the two programs are selected based on performance. GP algorithm copies the two winner programs and transforms these copies into two new programs via crossover and mutation as per fitness [17]. Creation of offsprings continues iteratively till a specified number of offsprings are created in the generation, and until a specified number of generations are produced. The resulting offspring at the end of the process is either a computer program or an equation. It is the solution to the problem. The generated equation or program can be directly applied to unseen data to obtain required predictions [25].

Figure 6.2 shows a typical GP flowchart. GP can evolve a formula or an equation by automatically selecting the significant inputs contributing to the model development and simultaneously ignoring the variables with the least contribution to the model. Three forms of GP are available, namely standard or tree-based GP, linear GP (LGP), and automatic induction of machine (AIM) code GP. Tree-type GP operates on parse trees, which is constituted by a function set and a terminal set. Crossover operation is carried out by randomly swapping the subtrees between the selected individuals [2]. The typical GP process is continued until the stopping criterion is met. LGP evolves sequences of instructions using an “imperative programming language” or a machine [8]. In LGP, the expressions of a functional programming language (like LISP) are substituted by programs of imperative language (like C/C++) [3]. Noneffective codes (introns) are identified and eliminated. It allows the corresponding useful instructions to be extracted from a program, and the process can be accelerated [7]. AIMGP is a particular form of LGP. For the given dataset, AIMGP writes a computer program and results with the “Best Program” as well as “Best Team.” It creates a total of 30 programs in a project and produces five teams of 1,3,5,7,9 program combinations [5].

6.5 Case study in India

In the present case study, one day and two days ahead air pollutant forecasting models have been developed for Pune city (18.5218.5204°N, 73.8567°E), which is one of the highly polluted metropolitan cities located in the Maharashtra state of India. It is the premier industrial center and the fastest developing city of the country [37]. The city is located on the western side of the Maharashtra state on the Deccan plateau at an elevation of 560 m above mean sea level at Karachi. Rapid industrialization along with boom in construction sector has resulted in exponential vehicular population growth, which has resulted in increased pollutant concentrations and worsening of the air quality of Pune for the last few years [32]. Central Pollution Control Board (CPCB) of India has declared SO_2 , NO_x , and RSPM as the criteria air pollutants which decide the air quality [35]. The national ambient air quality standards of India sets the maximum permissible limit for daily average concentrations of SO_2 , NO_x , and RSPM as 80, 80, 100 $\mu\text{g}/\text{cum}$, respectively, for residential urban areas. For the last decade, it has been observed that the upper limits have been crossed for SO_2 , NO_x , and RSPM with a maximum value recorded as high as 87,158 and 680 $\mu\text{g}/\text{cum}$, respectively, for the study area [33].

Daily values of the pollutant concentrations recorded by Maharashtra pollution control board from January 2005 to July 2015 were used for the development of annual and seasonal models. In order to obtain the complete annual picture about the air quality of Pune, the annual models are developed by arranging the yearly data

Figure 6.2: Typical GP Flowchart.
Source: Yadav & Sharma, 2019.

Figure 6.3: Time series plot for 1 day ahead temporal annual SO_2 .

Figure 6.4: Time series plot for 2 days ahead temporal annual SO₂.

Figure 6.5: Time series plot for 1 day ahead temporal annual NO_x .

Figure 6.6: Time series plot for 2 days ahead temporal annual NO_x

Figure 6.7: Time series plot for 1 day ahead temporal annual RSPM.

Figure 6.8: Time series plot for 2 days ahead temporal annual RSPM.

serially [16]. Similarly, pollutant concentrations vary with the environmental conditions as observed in different seasons. In India, four seasons such as winter (Jan–Feb), summer (Mar–May), monsoon (Jun–Sept), and post monsoon (Oct–Dec) are observed [24]. Seasonal models are also developed by dividing the data into different seasons.

6.5.1 Experimentation using ANN and GP

Air quality is time-dependent phenomenon. In this chapter, using soft computing approaches of ANN and GP, annual and seasonal models have been developed with the time series data of the criteria air pollutant concentrations. Uniform seasonal patterns and linearity between the inputs were assumed while developing all the forecasting models [26]. The data pertaining to the pollutant concentrations (SO_2 , NO_x , and RSPM) recorded from January 2005 to July 2015 have been considered for model development. The input optimization has been carried out using correlation analysis, considering the annual temporal data spanning over the week. Inputs were added one by one, and the multiple linear regression (MLR) models were developed and evaluated. The correlation coefficient of the MLR models suggested that three, four, and three predecessor values shall be used as model inputs for SO_2 , NO_x , and RSPM models, respectively. The same number of inputs was used for respective seasonal models as well.

Temporal annual 1 day ahead forecasting models can be written as [29]

$$\text{SO}_2(t+1) = f(\text{SO}_2(t), \text{SO}_2(t-1), \text{SO}_2(t-2)) \quad (6.1)$$

$$\text{NO}_x(t+1) = f(\text{NO}_x(t), \text{NO}_x(t-1), \text{NO}_x(t-2), \text{NO}_x(t-3)) \quad (6.2)$$

$$\text{RSPM}(t+1) = f(\text{RSPM}(t), \text{RSPM}(t-1), \text{RSPM}(t-2)) \quad (6.3)$$

Temporal annual 2 days ahead forecasting models can be written as

$$\text{SO}_2(t+2) = f(\text{SO}_2(t), \text{SO}_2(t-1), \text{SO}_2(t-2)) \quad (6.4)$$

$$\text{NO}_x(t+2) = f(\text{NO}_x(t), \text{NO}_x(t-1), \text{NO}_x(t-2), \text{NO}_x(t-3)) \quad (6.5)$$

$$\text{RSPM}(t+2) = f(\text{RSPM}(t), \text{RSPM}(t-1), \text{RSPM}(t-2)) \quad (6.6)$$

and temporal seasonal 1 day ahead forecasting models can be written as

$$\begin{aligned} \text{SO}_2(t+1) &= f(\text{SO}_2(t), \text{SO}_2(t-1), \text{SO}_2(t-2)) \\ &\text{for winter/summer/monsoon/post monsoon} \end{aligned} \quad (6.7)$$

$$\begin{aligned} \text{NO}_x(t+1) &= f(\text{NO}_x(t), \text{NO}_x(t-1), \text{NO}_x(t-2), \text{NO}_x(t-3)) \\ &\text{for winter/summer/monsoon/postmonsoon} \end{aligned} \quad (6.8)$$

$$\text{RSPM}(t+1) = f(\text{RSPM}(t), \text{RSPM}(t-1), \text{RSPM}(t-2)) \quad (6.9)$$

for winter/summer/monsoon/postmonsoon

Similarly, temporal seasonal 2 days ahead forecasting models can be written as

$$\text{SO}_2(t+2) = f(\text{SO}_2(t), \text{SO}_2(t-1), \text{SO}_2(t-2)) \quad (6.10)$$

for winter/summer/monsoon/postmonsoon

$$\text{NO}_x(t+2) = f(\text{NO}_x(t), \text{NO}_x(t-1), \text{NO}_x(t-2), \text{NO}_x(t-3)) \quad (6.11)$$

for winter/summer/monsoon/postmonsoon

$$\text{RSPM}(t+2) = f(\text{RSPM}(t), \text{RSPM}(t-1), \text{RSPM}(t-2)) \quad (6.12)$$

for winter/summer/monsoon/postmonsoon

One day and 2 days ahead, pollutants are predicted using ANN and GP with control parameters listed in Tables 6.1 and 6.2.

Table 6.1: Control parameters used in ANN.

S. no.	Items	Criteria used in this chapter
1	Network architecture	Input neurons = 3, 4, and 3 for SO_2 , NO_x , and RSPM Output neurons = 1 variable for each model Hidden neurons = smallest number of neuron which yields minimum prediction error on the validation data set
2	Neuron activation function	Input neuron = identity function Output neuron = identity function Hidden neuron = hyperbolic tangent function logsig and purelin for all the models
3	Learning parameters	The learning parameters converge to the network configuration and give best performance on validation data set with least epochs
4	Criteria for initialization of the initial weights	Network weights are uniformly distributed in the range of -1 to 1
5	Training algorithm	Levenberg Marquardt
6	Stopping criteria	Performance goal/epochs
7	Performance indicator	<i>R</i>, RMSE, <i>d</i>, MBE

Table 6.2: Control parameters used in GP.

S. no.	Parameters used	Value
1	Maximum initial tree size	15–20
2	Maximum tree size	45–50
3	Population size	50–1,000
4	No. of children produced	100–3,000
5	Mutation	0.05
6	Crossover rate	0.4–1.00
7	Objective type	R^2 , RMSE

6.6 Results and discussion

We have demonstrated the application of GP, which is one of the evolutionary algorithms, a subset of machine learning to forecast the air quality parameters a few time steps in advance. One day and two days ahead, temporal annual and temporal seasonal models are developed and evaluated using ANN and GP. Owing to the numerous advantages, ANN has been commonly used for air quality forecasting [30]. It has the capability of adaptive learning self-organization and handling of the nonlinear systems [20]. We have used ANN as a benchmarking tool in this work. GP is relatively a new approach that has been used for air quality forecasting of Pune city. ANN and GP models have been developed and tested for unseen inputs, and the qualitative and quantitative performance is assessed by the correlation coefficient (r), root mean squared error (RMSE), d -statistics (d), and mean bias error (MBE).

6.6.1 Annual models

Temporal one day and two days ahead forecasting of air quality parameters has been carried out using ANN and GP.

6.6.1.1 Forecasting of SO₂

Sulfur dioxide is an air pollutant that is responsible for acid rain, haze, and many health-related problems. It is produced when the coal is burnt to generate electricity. As per CPCB report, 2010, the chief source of SO₂ in Pune city is industry

(73.53%) followed by mobile sources (12.70%) and brick kilns (6%). Industrial area in Pune has a very high fuel consumption. It is the largest forging industry in Asia. Observed concentrations of SO_2 are slightly higher than the permissible limits. Both the soft computing approaches, namely, ANN and GP worked reasonably well as evident from the performance indicators, but the performance of GP is slightly better than ANN for the present case study (refer Table 6.3). In the process of predictions using ANN, the structure is defined initially, and the learning algorithm finds the weights. In GP, functions are defined, which always results in an optimal solution. One day ahead prediction results are better than two days ahead as it can be observed from Fig. 6.3 and Fig. 6.4 as it can be seen with increased r .

Correlation coefficient of $\text{SO}_2(t+1)$ for GP is 0.65 and for $\text{SO}_2(t+2)$ is 0.6 and d -statistic of $\text{SO}_2(t+1)$ for GP 0.8 and for $\text{SO}_2(t+2)$ for GP is 0.75. The peak observed in testing for one day ahead prediction was $50 \mu\text{g}/\text{cum}$ and predicted by the ANN model is $44.35 \mu\text{g}/\text{cum}$. The peak predicted by GP is $45.60 \mu\text{g}/\text{cum}$. The observed peak for two days ahead forecasting was $46 \mu\text{g}/\text{cum}$. Peak prediction by ANN is $33.81 \mu\text{g}/\text{cum}$, and by GP it was $38.14 \mu\text{g}/\text{cum}$.

6.6.1.2 Forecasting of NO_x

NO_x is produced in the air from the reaction of nitrogen and oxygen gases, especially at high temperatures, during combustion. The amount of nitrogen oxides emitted into the atmosphere as air pollutant can be significant, especially in the areas of high motor vehicle traffic in Pune city. Increased levels of nitrogen oxides can cause damage to the human respiratory tract and increase the risk of respiratory infections and asthma. Chronic lung diseases can be a result of long-term exposure to high levels of nitrogen dioxides. According to the CPCB report 2010, the significant contribution of 95% of NO_x in Pune is from vehicles. The industrial contribution is limited to 2% as industrial area is confined within the city limits. Remaining 3% of NO_x is produced from domestic and commercial fuel burning for cooking. Modeling and forecasting of NO_x can help to warn the residents, especially sensitive groups. It was observed that the concentrations of NO_x exceeded the annual standards during study period.

One day and two days ahead NO_x forecasting models have been developed and evaluated using ANN and GP. Alike SO_2 , it can be seen that GP results are better than ANN and one day ahead predictions are superior than two days ahead (Tables 6.3 and 6.4 and also from Fig. 6.5 and Fig. 6.6). It can be seen that NO_x forecasts are more accurate than SO_2 forecasts. As far as errors are concerned, RMSE values are higher than as compared to SO_2 models, and negative MBE is observed for NO_x models indicating that the models are under predicting.

The peak observed in testing for one day ahead prediction was $115 \mu\text{g}/\text{cum}$ and predicted by the ANN model is $53.64 \mu\text{g}/\text{cum}$. The peak predicted by GP is

Table 6.3: One day ahead temporal annual model results.

Performance evaluator	Models	Pollutant SO ₂	NO _x	RSPM
<i>r</i>	ANN	0.6673	0.8110	0.8374
	GP	0.6776	0.8258	0.8434
RMSE	ANN	6.1876	11.6846	32.4299
	GP	6.0417	12.2770	31.6747
<i>d</i>	ANN	0.7989	0.8863	0.8956
	GP	0.7995	0.8943	0.9088
MBE	ANN	0.5077	0.2663	-2.1348
	GP	0.2041	-0.0391	-0.5727

Table 6.4: Two days ahead temporal annual model results.

Performance evaluator	Models	Pollutant SO ₂	NO _x	RSPM
<i>r</i>	ANN	0.6523	0.8109	0.7006
	GP	0.6487	0.7752	0.7864
RMSE	ANN	6.2901	11.6856	42.3596
	GP	6.2298	12.6171	36.4318
<i>d</i>	ANN	0.7880	0.8862	0.8173
	GP	0.7585	0.8591	0.8713
MBE	ANN	0.4669	-0.2707	-2.1363
	GP	0.3714	-0.1105	-1.2505

123.45 µg/cum. Similarly, the observed peak for two days ahead forecasting was 155 µg/cum, and peak predicted by ANN and GP is 80.26 and 83.83 µg/cum, respectively, which indicates that GP predicts the peak closer to the observed values.

6.6.1.3 Forecasting of RSPM

Particulate matters with an aerodynamic diameter less than or equal to $10\text{ }\mu\text{m}$ are classified as respirable suspended particulates (RSP). Particulates are produced from combustion processes, vehicles, and industrial sources. They are microscopically small and can enter the lungs and penetrate deeply into the respiratory system and cause difficulty in breathing as well as respiratory and cardiovascular effects such as bronchitis and asthma. Sensitive groups, including pregnant woman, children, and older people, are at high risk. A warmer climate increases the risk of illnesses and death from extreme heat and poor air quality. Pune city has a hot and semiarid climate, which further increases the risk [32]. Vehicular emission is a major source of particulate matter in Pune and the highest recorded concentration of RSPM is $680\text{ }\mu\text{g}/\text{cum}$ during the study period. RSPM is the pollutant responsible for deciding the air quality. Modeling and forecasting of RSPM are essential to calculate the air quality index, which can be easily understood by the community.

From the observation table 6.3 and 6.4 and also from Fig. 6.7 and Fig. 6.8, it can be seen that similar to SO_2 and NO_x results, one day ahead forecasts of RSPM are better than two days ahead, and the GP models performed better than ANN. The fact that air pollution is a time-dependent phenomenon is verified from the results. There is a strong correlation between previous concentrations of RSPM with the forecasted values, which is evident from r values indicating a stronger correlation. “ r ” is a linear measure, and the air quality model evaluations based on only “ r ” mostly fail due to the presence of lag between source emission quantity and the ambient pollutant concentration. Temporal models fail to take into account the lag effect as the meteorological parameters are absent as inputs. The lag is observed due to inversion condition, which indicates the accumulation of pollutants in the ambient environment during odd hours of the day when there are no source emissions recorded [13]. In such conditions for air quality modeling, “ d ” statistics is the relevant evaluation criteria. “ d ” values of 0.90 and 0.87 are recorded for RSPM one day and two days ahead forecasting, which indicates that about 90% and 87% model results are error free. Negative MBE demonstrates that both the tools, ANN and GP, underpredict the phenomenon. Lower RMSE values ($31.67\text{ }\mu\text{g}/\text{cum}$ for one day and $36.43\text{ }\mu\text{g}/\text{cum}$ for two days ahead forecasting) in the data values ranging between 17 and $680\text{ }\mu\text{g}/\text{cum}$ indicate the satisfactory performance of GP.

Alike SO_2 and NO_x models, peak predicted by GP for RSPM one day ahead model is closer to the actual value (observed peak is $291\text{ }\mu\text{g}/\text{cum}$ and the peak predicted by ANN is $242.55\text{ }\mu\text{g}/\text{cum}$ and by GP is $244.12\text{ }\mu\text{g}/\text{cum}$), whereas the observed testing peak for two days ahead prediction was 680, but the peak predicted by ANN is $145.36\text{ }\mu\text{g}/\text{cum}$ and by GP is $153.28\text{ }\mu\text{g}/\text{cum}$. This could be due to the fact that RSPM concentration of $680\text{ }\mu\text{g}/\text{cum}$ was observed only once during Diwali (festival) day.

6.6.2 Seasonal models

Seasonality is one of the essential factors affecting air quality. Variations in the pollutant concentrations are observed in different seasons due to the interaction of the various meteorological parameters with pollutant concentrations [27]. Seasonal models are formulated and evaluated using ANN and GP to get the complete picture of the air quality of Pune.

In Pune city, winter is observed during January and February. Air quality gets worsened during this time due to inversion condition. In winter time, a temperature inversion takes place when a layer of warmer air traps cold air near the ground. During inversion, the air becomes stagnant and pollutants are caught close to the land, which may result in the formation of smog.

Summer starts at the beginning of March and ends at the end of May. As compared to winter, summer heat prevents the inversion, which improves the air quality slightly. Summer always seems to drive out the dense clouds of pollution that suffocate many Indian citizens. Air pollutant concentrations are observed to be higher during the summer time because the heat and sunlight primarily react with the air along with all the chemical compounds in it. When these compounds react with the nitrogen oxide emissions present in the air, it results in smog. According to medical studies, air pollution in summer increases the risk of heart attack [31].

June to September is declared as monsoon months in Pune [26]. Rainfall is a cleansing mechanism that results in washing away of the pollutants during monsoon [26]. Rain makes the ground wet and mixing of the dust into the air again becomes difficult resulting in lesser pollutant concentration. Ozone formation rate is also lower in monsoon as the rain clouds obstruct the correct amount of temperature and sunlight required for ozone formation.

October and November are the two months of post monsoon (or retreating monsoon) in which the wind becomes weaker. Skies are clear, but humidity in the atmosphere increases. Diwali festival happens to be in this season. The air quality in this season decreases due to firecrackers. This condition increases the chance of pollution episodes.

All four seasons have different characteristics and from the observation tables, it can be seen that both the tools (ANN and GP) worked well, but GP results are better than ANN.

From the observation Tables 6.5–6.7, it can be seen that one day ahead predictions are superior to two days ahead forecasts. Results indicate that SO_2 models have shown better performance in summer and post-monsoon season. NO_x models have performed well in monsoon and post-monsoon and all RSPM models performed very well in all the seasons except for winter. The probable reason could be in winter there are slight movements of the air masses resulting in stagnation of the pollutants.

It was observed that maximum concentrations of SO_2 , NO_x , and RSPM were recorded in winter, whereas the minimum concentrations have been recorded in

Table 6.5: Seasonal forecasting of SO₂.

Pollutant	Time of prediction	Season	Tool	Performance evaluator			
				<i>r</i>	RMSE	<i>d</i>	MBE
SO ₂	1 day ahead	Winter	ANN	0.4734	8.4711	0.6688	0.7368
			GP	0.5577	7.5759	0.6963	0.3866
		Summer	ANN	0.6544	4.8672	0.7408	-0.0325
			GP	0.6546	4.7015	0.8267	0.3179
		Monsoon	ANN	0.5692	5.1757	0.7408	-0.0325
			GP	0.7067	4.3710	0.8267	0.8267
		Post monsoon	ANN	0.6338	7.8940	0.7865	0.6965
			GP	0.7066	6.8280	0.8219	0.3703
	2 days ahead	Winter	ANN	0.3024	9.0665	0.5142	1.0553
			GP	0.3588	8.6694	0.5503	-0.0517
		Summer	ANN	0.6022	5.4609	0.6044	2.0816
			GP	0.5914	5.0841	0.7040	0.9720
		Monsoon	ANN	0.5556	5.3268	0.7079	0.0609
			GP	0.5734	5.0645	0.7919	0.1971
		Post monsoon	ANN	0.4560	8.7618	0.6392	0.4464
			GP	0.4582	8.6370	0.6161	0.4400

monsoon as well as winter. Heavy fog during early mornings of winter supports the highest concentrations during winter. Rainfall was observed during study time in winter for few days resulting in the minimum pollutant concentrations. Rainfall acts as a cleansing agent and results in washing away of the pollutants in monsoon. From the meteorological records, it can be seen that the quantity of rainfall was sufficient to reduce the concentrations of the contaminants in monsoon.

6.7 Modeling options for situations when there are constraints on the data availability

Quality and consistency of the data decides the model prediction accuracy. Availability of continuous and accurate meteorological and pollutant concentration

Table 6.6: Seasonal forecasting of NO_x .

Pollutant	Time of prediction	Season	Tool	Performance evaluator			
				<i>r</i>	RMSE	<i>d</i>	MBE
NO_x	1 day ahead	Winter	ANN	0.6495	17.2756	0.7647	-1.1737
			GP	0.6846	16.4143	0.7849	-1.4931
		Summer	ANN	0.6758	19.3564	0.8681	-0.4552
			GP	0.6920	18.7333	0.8810	-0.1921
		Monsoon	ANN	0.7221	8.1978	0.8163	0.5864
			GP	0.7784	7.4318	0.8687	0.3330
		Post monsoon	ANN	0.7248	13.1202	0.8074	-1.7946
			GP	0.8080	11.0645	0.8848	0.0178
	2 days ahead	Winter	ANN	0.5625	18.6040	0.6885	-1.5606
			GP	0.6511	17.0322	0.7722	-0.6801
		Summer	ANN	0.5860	8.9691	0.7328	1.1892
			GP	0.5946	8.7939	0.7301	0.7584
		Monsoon	ANN	0.6352	9.4085	0.7155	2.0252
			GP	0.6775	8.7317	0.7903	0.5660
		Post monsoon	ANN	0.6245	14.9824	0.7135	-1.4086
			GP	0.6343	14.4177	0.7526	-0.0676

data is a challenging task in developing countries. Adverse meteorological conditions, instrumental errors, and so on are also some of the reasons for nonavailability of continuous and accurate data of all the meteorological parameters and pollutant concentrations. Atmospheric inversion is a vital condition deciding the air quality, which must be taken into account while modeling the air quality parameters. Data of meteorological parameters such as wind speed, rainfall, solar radiation, wind direction, humidity, mixing height, visibility, and color cover should be included as model inputs. Processing time for data collection and dissemination is quite high; as a result, it becomes difficult to include meteorology as a model input as in the case of this work. In such situations, the models developed should be robust enough so that they can be useful for short-term prediction of criteria air pollutants. An attempt has been made to suggest various air quality modeling alternatives.

Table 6.7: Seasonal forecasting of RSPM.

Pollutant	Time of prediction	Season	Tool	Performance evaluator			
				<i>r</i>	RMSE	<i>d</i>	MBE
RSPM	1 day ahead	Winter	ANN	0.5244	30.8937	0.6563	9.4319
			GP	0.5553	28.8443	0.6839	3.1098
		Summer	ANN	0.7935	19.3564	0.8681	-0.4552
			GP	0.8850	18.7333	0.8810	-0.1921
		Monsoon	ANN	0.7806	19.3564	0.8681	-0.4552
			GP	0.7964	18.7333	0.8810	-0.1921
		Post monsoon	ANN	0.6675	34.4305	0.7947	-0.5861
			GP	0.6826	33.6118	0.7914	-0.3259
	2 days ahead	Winter	ANN	0.3941	34.4683	0.8810	11.6724
			GP	0.3801	32.5242	0.8824	6.3909
		Summer	ANN	0.7729	52.3487	0.7881	-9.3464
			GP	0.7926	47.1344	0.8596	-3.0583
		Monsoon	ANN	0.6942	22.2994	0.8072	-0.5772
			GP	0.7104	21.7756	0.8122	-0.4177
		Post monsoon	ANN	0.4932	81.7014	0.4789	-31.2984
			GP	0.5527	14.4177	0.6815	-0.02559

When meteorological parameters are available, the correlation of inputs (meteorological parameters) with the output (pollutant concentrations) can be calculated and data can be arranged in the descending order of their correlation with the output. Only three inputs (from the available inputs) with high correlation with the output shall be considered and called as top three causes. Those having low correlation with the output are called as bottom three causes. The modeling options can be presented as [29]

Model type 1 – model with top three inputs (6.13)

Model type 2 – model with bottom three inputs (6.14)

Model type 3 – model with bottom three and top one input (6.15)

Model type 4 – temporal models (6.16)

Of the various meteorological parameters listed above, the temperature difference is the crucial factor for atmospheric inversion and, at the same time, it is effortless to record and disseminate. Temperature model can be developed in which previous values of the temperature difference can be added on the basis of correlation along with previous pollutant concentrations in case of the temporal model, and such modified temporal model can be presented as

$$\text{Model type 5 – temperature models} \quad (6.17)$$

With these modeling options, criteria air pollutants can be predicted on a continuous basis.

6.8 Conclusion and future directions

We have presented the application of soft computing techniques to forecast air quality parameters one and two days in advance. Models are developed with GP, and the results are compared with established forecasting tool of ANN. Temporal annual and seasonal models are developed for all the three criteria air pollutants (SO_2 , NO_x , and RSPM) and found that model results are matching with the recorded values. One day ahead forecasting results are better than two days ahead. GP proved to be better than ANN in terms of accuracy of forecast and found equally competent for peak predictions for this case study.

In air quality modeling, researcher always prefers a short-term and continuous model to broadcast the current status of the air quality. This information is useful to plan the activities and safeguard the health of sensitive groups. In developing countries like India, there is a problem of availability of continuous data. In such situations, various modeling options are proposed.

The major limitation of the study is the absence of meteorological parameters for model development. Model inputs were selected using a linear method of correlation for the air quality forecasting problem, which is actually a non-linear phenomenon. There were certain assumptions made while developing forecasting models such as seasonal patterns were assumed to be uniform, meteorological processes were simplified, and linearity between the inputs was assumed which is seldom true. The studies in future could be directed toward the better understanding of the tool GP and combining it with other air quality forecasting approaches. Comprehensive models can be prepared by considering the vertical and spatial distribution of the pollutants along with meteorological parameters.

References

- [1] Ahmad K., Khare M., & Chaudhry K.K. (2005) Wind tunnel simulation studies on dispersion at urban street canyons and intersections –a review, *Journal of Wind Engineering and Industrial Aerodynamics*, 93, 697–717.
- [2] Babovic V. & Keijzer M. (2000) Genetic programming as a model induction engine, *Journal of Hydroinformatics*, 2, 35–60.
- [3] Brameier M. & Banzhaf W. (2001) A Comparison of linear genetic programming and neural networks in medical data mining, *IEEE Transactions of Evolutionary Computations*, 5(1), 17–26.
- [4] Dixit P. & Londhe S.N. (2016) Prediction of Extreme wave heights using neuro wavelet techniques, *Applied Ocean Research*, 58, 241–252.
- [5] Francone F.D. & Deschaine L.M. (2004) Extending the boundaries of design optimisation by integrating fast optimization techniques with machine code based linear genetic programming, *Information Sciences*, 161(3–4), 99–120.
- [6] Gardner M.W. & Dorling S.R. (1998) Artificial neural networks (The Multilayer Perceptron)- A review of applications in the atmospheric sciences, *Atmospheric Environment*, 2627–2636.
- [7] Gandomi A.H., Alavi A.H., & Yun G.J. (2011(b)) Nonlinear modelling of shear strength of SFRC beams using linear genetic programming, *Structural Engineering Mechanics Techno press*, 38(1), 1–25.
- [8] Guven A. (2009) Linear Genetic Programming for time series modelling of daily flow rate, *Journal of Earth System Sciences*, 118(2), 137–146.
- [9] Jakeman A.J., Simpson R.W., & Taylor J.A. (1988) Modeling distributions of air pollutant concentrations – III, The hybrid deterministic-statistical distribution approach. *Atmospheric Environment*, 22(1), 163–174.
- [10] Londhe S.N. & Shah S. (2017) A Novel approach for knowledge extraction from artificial neural networks, *ISH Journal of Hydraulic Engineering*.
- [11] Bai L., Wang J., Xuejio M., & Haiyan L. (2018) Air pollution forecasts: An overview, *International Journal of Environmental Research and Public Health*, 15(780), 01–44.
- [12] Mishra D. & Goyal P. (2016) Neuro fuzzy approach to forecast NO₂ Pollutants addressed to air quality dispersion model over Delhi, India, *Aerosol and Air Quality Research*, 16, 166–174.
- [13] Nagendra S.M.S. & Khare M. (2006) Artificial neural network approach for modelling nitrogen dioxide dispersion from vehicular exhaust emissions, *Ecological Modelling*, 190, 99–115.
- [14] Niska H., Hiltunen T., Karppinen A., Ruuskanen J., & Kolehmainen M. (2004) Evolving the neural network model for forecasting air pollution time series, *Engineering Applications of Artificial Intelligence*, 17(2), 159–167.
- [15] The ASCE task Committee. (2000) Artificial neural networks in hydrology. I: Preliminary concepts, *Journal of Hydrologic Engineering*, 115–123.
- [16] Tikhe S.S., Khare K.C., & Londhe S.N. (2015) Multicity seasonal air quality index forecasting using soft computing techniques, *Advances in Environmental Research*, 4(02), 83–104.
- [17] Yadav A. & Sharma D. Enhancement of performance 32 bit RISC processor using Genetic Algorithms, *IJRECE*, 7(3), 449–454.
- [18] Lin Y., Zhao L., Haiyan L., & Sun Y. (2018) Air quality forecasting based on cloud model granulation, *EURASIP Journal on Wireless Communications and Networking*, 1–10.
- [19] Zhang Y., Bocquet M., Mallet V., Seigneur C., & Baklanov A. (2012) Real-time air quality forecasting, part I: History, techniques, and current status, *Atmospheric Environment*, 60, 632–655.

Books

- [20] Bose N.K. & Liang P. Neural network fundamentals with graphs, algorithms and applications. Tata McGraw Hill Publishing Company Limited, New Delhi, India, 1996.
- [21] Cheremisinoff N.P. & Paul N. Cheremisinoff. Pumps/Compressors/Fans Pocket Handbook Lancaster PA. Technomic Publishing Company, 1989.
- [22] Giorgio F. & Piero F. Mathematical models for planning and controlling air quality. Proceedings of IASA workshop 17, 1996.
- [23] Khare M. & Nagendra S.M.S. Artificial neural networks in vehicular pollution modelling. Springer, New York, 2007.
- [24] Koza J.R. Genetic programming. MIT Press, 1992.
- [25] Londhe S.N. & Dixit P. Genetic programming: A novel computing approach in modelling water flows. 2012.
- [26] Londhe S.N. & Panchang V. (2018) ANN techniques: A survey of coastal applications, Advances in Coastal Hydraulics.
- [27] Murthy P.B. Environmental Meteorology. I.K. International Publishing House Pvt. Ltd., New Delhi, 2009.
- [28] Rumelhart D.E. & James M. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press, Cambridge, 1986.
- [29] Tikhe Shruti K.K.C. & Londhe S.N. Forecasting of Air Quality Parameters using Soft Computing Techniques: Ph. D. thesis. Savitribai Phule Pune University, Pune, 2016.
- [30] Weber E. Air pollution, assessment methodology and modeling. Plenum, New York, 1982.

Websites

- [31] Air pollution in summer increases heart attack risks, 2019. (Accessed June,23, 2019 at <https://www.iqair.com/intl/node/811>.)
- [32] Air quality monitoring and emission source apportionment study for Pune, 2010. (Accessed June, 12, 2019 at http://mpcb.gov.in/ereports/pdf/pune_report_cpcb.pdf.)
- [33] Ambient air quality monitoring network in Maharashtra, 2016. (Accessed January, 15, 2016, at <http://mpcb.gov.in/>.)
- [34] How air pollution is destroying our health, 2020. (Accessed July 30, 2019 at <https://www.who.int/air-pollution/news-and-events/how-air-pollution-is-destroying-our-health>.)
- [35] India map, 2019 (Accessed June 23, 2019 at <https://www.mapsofindia.com/>.)
- [36] Overview of The clean air act and air pollution, 2016. (Accessed January, 10, 2016 at <https://www.epa.gov/clean-air-act-overview>.)
- [37] Pune air poses health threat to citizens, 2015. (Accessed December 4, 2015, at <https://www.dnaindia.com/pune/report-pune-air-poses-health-threat-to-citizens-2141315>.)