

Fuzzy-GA modeling in air quality assessment

Jyoti Yadav · Vilas Kharat · Ashok Deshpande

Received: 9 September 2014 / Accepted: 9 February 2015 / Published online: 14 March 2015
© Springer International Publishing Switzerland 2015

Abstract In this paper, the authors have suggested and implemented the defined soft computing methods in air quality classification with case studies. The first study relates to the application of Fuzzy C mean (FCM) clustering method in estimating pollution status in cities of Maharashtra State, India. In this study, the computation of weighting factor using a new concept of reference group is successfully demonstrated. The authors have also investigated the efficacy of fuzzy set theoretic approach in combination with genetic algorithm in straightway describing air quality in linguistic terms with linguistic degree of certainty attached to each description using Zadeh–Deshpande (ZD) approach. Two metropolitan cities viz., Mumbai in India and New York in the USA are identified for the assessment of the pollution status due to their somewhat similar geographical features. The case studies infer that the fuzzy sets drawn on the basis of expert knowledge base for the

criteria pollutants are not much different from those obtained using genetic algorithm. Pollution forecast using various methods including fuzzy time series forms an integral part of the paper.

Keywords Genetic algorithm · Linguistic description · Fuzzy logic, fuzzy inference system · ZD approach · Fuzzy C mean · Clustering · Fuzzy time series

Introduction

The composition of the atmosphere has been gradually changing over the past millions of years. Rapid urbanization and industrialization has added other elements/compounds to the pure air and thus caused the increase in pollution. In addition to the criteria pollutants, there are many potent short-lived climate pollutants like black carbon, tropospheric ozone, methane, and hydrofluorocarbons that need to be regulated to reduce emissions. These pollutants remain in the atmosphere for shorter periods of time and have much larger global warming potentials compared to CO₂. As one of the most significant sources of greenhouse gases (GHG) and criteria pollutant emissions, the transportation system represents one of the greatest needs for emission reductions and one of the greatest opportunities to build an economy that aligns stable economic growth with the need for ever-improving public health and environmental protection. Reducing transportation emissions, including those from heavy-duty diesel engines, will have dramatic air quality and public health benefits. Black

J. Yadav (✉) · V. Kharat
Department of Computer Science, University of Pune, Pune,
India
e-mail: jyo_yadav@yahoo.co.in

V. Kharat
e-mail: laddool@yahoo.com

A. Deshpande
Berkeley Initiative Soft Computing (BISC)-Special Interest
Group (SIG) Environment Management System (EMS),
Berkeley, CA, USA
e-mail: ashok_deshpande@hotmail.com

A. Deshpande
College of Engineering Pune (COEP), Pune, India

carbon (as a component of $PM_{2.5}$) and ozone are the air pollutants with harmful health effects and reducing their emissions can offer significant improvements in air quality and public health. In addition to the short-lived, local ozone precursors like NO_x and SO_x , methane is a global source of tropospheric ozone. Black carbon is the most strongly light-absorbing component of particulate matter (PM) emitted from burning fuels such as coal, diesel, and biomass. Diesel PM is a toxic air contaminant that can be inhaled (PM_{10} and $PM_{2.5}$). Black carbon contributes to climate change both directly by absorbing sunlight and indirectly by depositing on snow and by interacting with clouds and affecting cloud formation. Reducing black carbon emissions globally can have immediate economic, climate, and public health benefits (Brown et al. 2006).

While we must continue taking steps to rapidly reduce CO_2 emissions, additional efforts to cut emissions of short-lived climate pollutants can yield immediate climate benefits. In addition, fast and sustainable actions to reduce these emissions can help to achieve other benefits though avoided impacts on agriculture, water availability, ecosystems, and human health. The reduction of methane would reduce background tropospheric ozone concentrations, which would help with progress towards healthy air quality and avoid crop yield losses and forest damage due to the direct action of ozone on plant growth. Black carbon impacts cloud formation and precipitation, and black carbon deposits on glaciers and snowpack accelerate melting. Reducing black carbon and methane emissions will help reduce the risk for premature deaths, air pollution-related hospitalizations, and associated medical expenses each year. With this backdrop, the need for enactment of pollution control laws was recognized by the US-EPA (Brown et al. 2006).

In order to prevent, control, and abate air pollution, the Clean Air Act was enacted in 1970 by US-EPA and Air Quality Index (AQI) for defining air quality was suggested. AQI formulation uses probability theory and is devoid of the concept of partial belief associated with human thinking. The uncertainty associated with criteria pollutants is considered as probabilistic.

The modeling formalism has embraced two valued logic-based probability theory wherein random variable is used as the basis of probability computations. The standard probability theory is not designed to deal with imprecise probabilities or Z-probabilities which pervade

real-world uncertainties. Fuzzy set theory is the way of modeling uncertainty due to imprecision, fuzziness, and ambiguity wherein human perception plays a pivotal role.

The limitations of conventional AQI calls for devising fuzzy logic-based formalism, known as *Zadeh–Deshpande (ZD)* approach (Yadav et al. 2011, 2013, 2014), wherein air quality/water quality is described straightway in linguistic terms with linguistic degree of certainty attached to each description.

The overall focus of the paper is centered on the linguistic description of air quality using genetic algorithm–fuzzy modeling in combination with ZD approach via case studies. The self-explanatory system flowchart in Fig. 1 shows the approaches used in the research study.

The remaining part of the paper is organized as follows. Section “**Classification of polluted cities of Maharashtra State in India using Fuzzy C mean**” describes classification of 15 cities in Maharashtra, India, based on their pollution potential using Fuzzy C mean (FCM). Section “**Pollution forecast and application of GA-ZD method**” predicts pollution forecast using four mathematical techniques and genetic algorithm–fuzzy modeling in combination with ZD approach to assess and compare the air quality status of New York and Mumbai. Section “**Software support**” gives details regarding the software developed for linguistic description of air quality. Concluding remarks and future scope for research are presented in “**Conclusion and future scope for research**” section.

Classification of polluted cities of Maharashtra State in India using Fuzzy C mean

Clustering of numerical data forms the basis of many classification and system modeling algorithms. The purpose of clustering is to identify natural groupings of data from a large data set to produce a concise representation of a system’s behavior. Clustering refers to identifying the number of subclasses of c clusters in a data universe X comprised of n data samples and partitioning X into c clusters ($2 \leq c < n$). Note that $c=1$ denotes rejection of the hypothesis that there are clusters in the data, whereas $c=n$ constitutes the trivial case where each sample is in a “cluster” by itself. There are two kinds of c partitions of

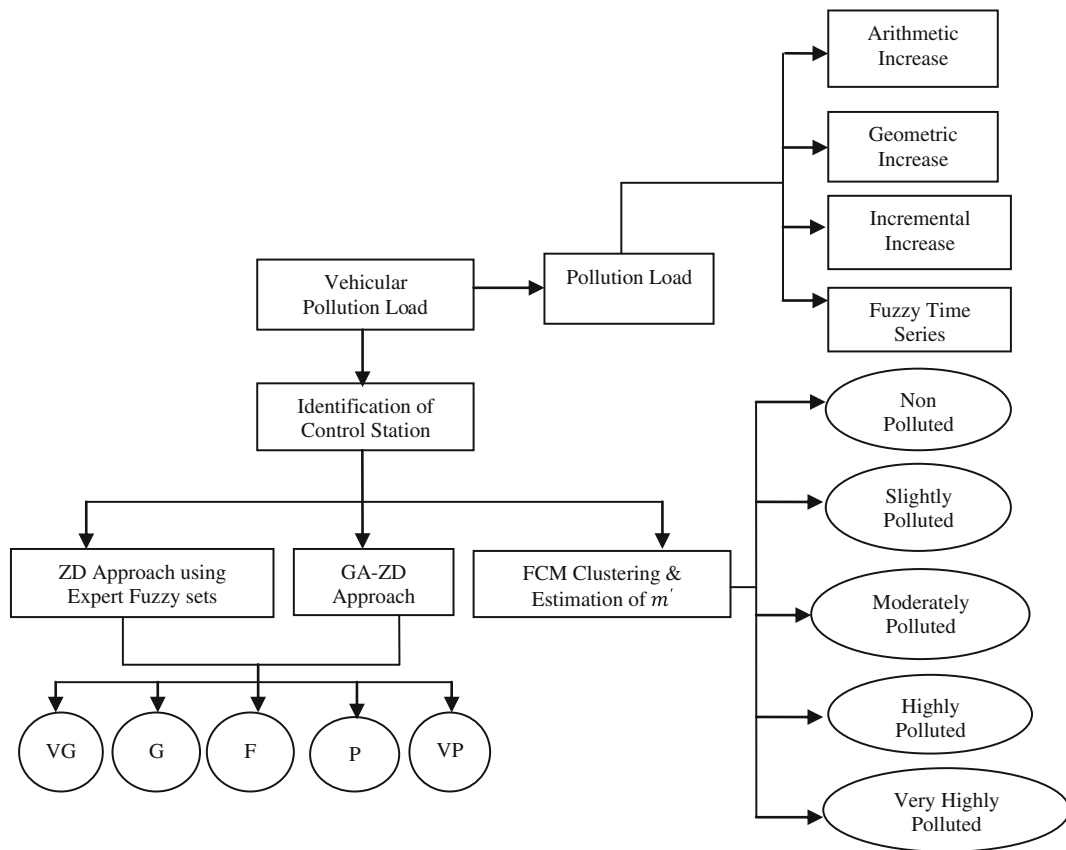


Fig. 1 System flowchart

data: hard (or crisp) and soft (or fuzzy). For numerical data, one assumes that the members of each cluster bear more mathematical similarity to each other than to members of other clusters. Two important issues to consider in this regard are how to measure the similarity between pairs of observations and how to evaluate the partitions once they are formed (Ross 2009).

One of the simplest similarity measures is distance between pairs of feature vectors in the feature space. If one can determine a suitable distance measure and compute the distance between all pairs of observations, then one may expect that the distance between points in the same cluster will be considerably less than the distance between points in different clusters. Several circumstances, however, mitigate the general utility of this approach, such as the combination of values of incompatible features, as would be the case, for example, when different features have significantly different scales. The clustering methods described

defines “optimum” partitions through a global criterion function that measures the extent to which candidate partitions optimize a weighted sum of squared errors between data points and cluster centers in feature space. Hard C means are employed to classify data in a crisp sense. By this, we mean that each data point will be assigned to one, and only one, data cluster. The problem of hard C mean lies in assigning the point on the line of symmetry to a class. To which class should this point belong? Whichever class the algorithm assigns this point to, there will be a good argument that it should be a member of the other class. Alternatively, the argument may revolve around the fact that the choice of two classes is a poor one for this problem. Three classes might be the best choice, but the physics underlying the data might be binary and two classes may be the only option (Ross 2009).

The solution to the above problem is FCM. This technique was originally introduced by Jim Bezdek in

1981 as an improvement on earlier clustering methods. FCM is a data clustering technique in which a dataset is grouped into n clusters with every data point in the dataset belonging to every cluster to a certain degree. For example, a certain data point that lies close to the center of a cluster will have a high degree of belonging or membership to that cluster and another data point that lies far away from the center of a cluster will have a low degree of belonging or membership to that cluster. Thus each data point belongs to a cluster to some degree that is specified by a membership grade (Bezdek et al. 1984; Saksena et al. 2003).

The case study relates to clustering of 15 cities of Maharashtra (India), based on their pollution potential, into five different clusters. The classification is based on the pollution levels of Ambient Air Quality Monitoring (AAQM) stations.

A sample set of n data samples that we wish to classify: $X = \{x_1, x_2, x_3, \dots, x_n\}$. Each data sample, x_i , is defined by m features, i.e., $x_i = \{x_{i1}, x_{i2}, x_{i3}, \dots, x_{im}\}$ where each x_j in the universe X is an m -dimensional vector of m elements or m features. There are 15 data points (cities), and each data point is described with two features viz. NO_x and PM_{10} . The 15 data points are to be clustered into five clusters viz. *not polluted*, *moderately polluted*, *poorly polluted (alert level)*, *very poorly polluted (warning level)*, and *severely polluted (emergency level)*. The matrix U has been arranged as follows:

$$U^0 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 11 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \end{bmatrix}$$

V_i is the i^{th} cluster center, which is described by m features and can be arranged in vector form as $V_i = \{v_{i1}, v_{i2}, v_{i3}, \dots, v_{im}\}$. Each of the cluster coordinates for each class can be calculated using Eq. (1)

$$v_{ij} = \frac{\sum_{k=1}^n \mu_{ik}^{m'} \cdot x_{kj}}{\sum_{k=1}^n \mu_{ik}^{m'}} \tag{1}$$

where j is a variable on the feature space, i.e., $j=1,2,3,\dots, m$, and μ_{ik} is the membership of the k^{th} data point in the i^{th} cluster. The distance measure d_{ik} is a Euclidean distance between the i^{th} cluster center and the k^{th} data set computed using Eq. (2). $m' \in [1, \infty)$ is the weighting parameter that controls the amount of fuzziness in the clustering process. In the present case study, $m'=2.2$.

$$d_{ik} = \sqrt{\sum_{j=1}^m (x_{kj} - v_{ij})^2} \tag{2}$$

with the distance measures update U using Eq. (3)

$$\mu_{ik}^{(r+1)} = \left[\sum_{j=1}^c \left(\frac{d_{ik}^{(r)}}{d_{jk}^{(r)}} \right)^2 \right]^{-1} \tag{3}$$

Thus the updated fuzzy partition obtained is:

$$U^1 = \begin{bmatrix} 0.39 & 0.38 & 0.42 & 0.23 & 0.37 & 0.22 & 0.41 & 0.40 & 0.06 & 0.04 & 0.38 & 0.160.66 & 0.25 & 0.40 \\ 0.50 & 0.61 & 0.56 & 0.31 & 0.47 & 0.29 & 0.55 & 0.52 & 0.12 & 0.06 & 0.49 & 0.240.33 & 0.34 & 0.51 \\ 0.04 & 0.01 & 0.01 & 0.15 & 0.05 & 0.17 & 0.01 & 0.03 & 0.15 & 0.24 & 0.05 & 0.150.00 & 0.13 & 0.03 \\ 0.05 & 0.01 & 0.01 & 0.19 & 0.07 & 0.21 & 0.02 & 0.03 & 0.25 & 0.40 & 0.06 & 0.210.00 & 0.17 & 0.04 \\ 0.04 & 0.01 & 0.01 & 0.18 & 0.07 & 0.18 & 0.02 & 0.03 & 0.48 & 0.27 & 0.06 & 0.320.00 & 0.17 & 0.04 \end{bmatrix}$$

To determine whether we have achieved convergence, we choose a matrix norm such that the maximum absolute value of pairwise comparisons of each of the values in U^0 and U^1 using Eq. (4).

$$\max_{i,k} \left| \mu_{ik}^{(1)} - \mu_{ik}^{(0)} \right| \tag{4}$$

Finally, the FCM converges after four iterations, and the following clusters are obtained (Fig. 2)

1. Cluster 1 = {Akola, Kolhapur, Ratnagiri, Solapur} → not polluted
2. Cluster 2 = {Amravati, Aurangabad, Jalgaon, Nashik, Sangli} → slightly polluted
3. Cluster 3 = {Pune} → moderately polluted
4. Cluster 4 = {Chandrapur, Jalna, Latur, Nagpur} → highly polluted
5. Cluster 5 = {Mumbai} → very highly polluted

Table 1 presents air quality description of the clusters by inference. Based on the available criteria pollutant

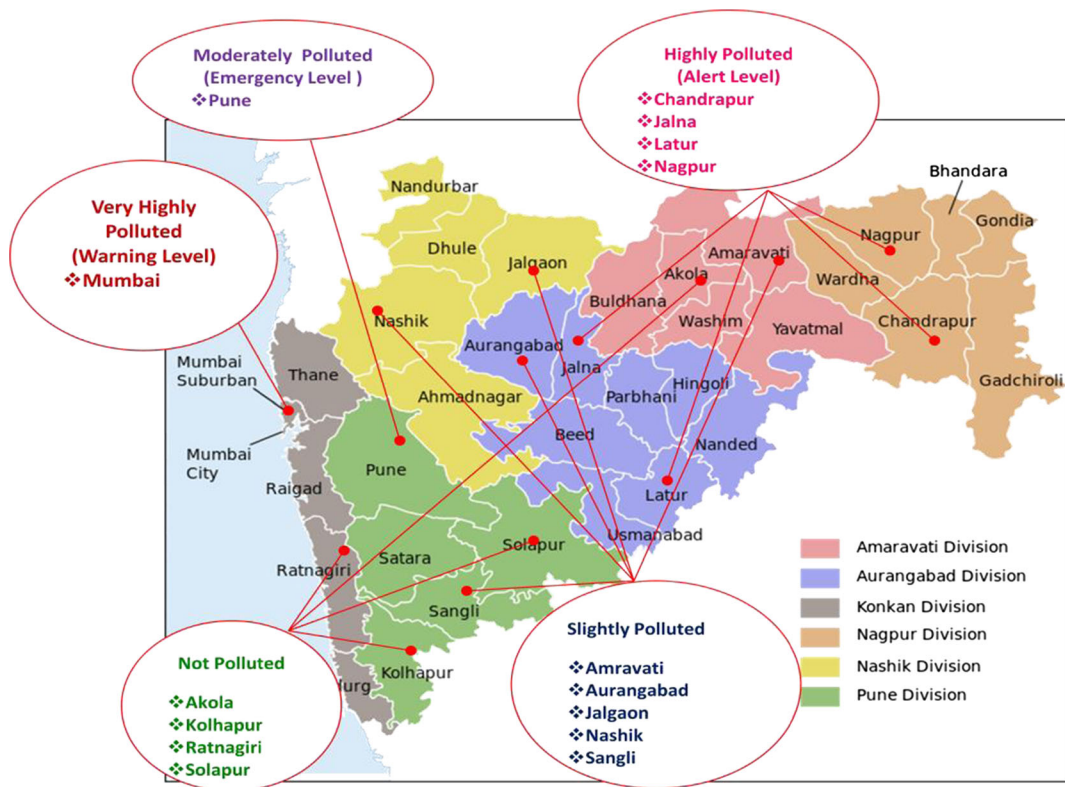


Fig. 2 FCM classification: pollution status of 15 cities in Maharashtra (India)

data for the year 2009, Ratnagiri is classified as *not polluted*. It lies at the heart of Konkan region, a charming stretch of land on the west coast of India, endowed with beautiful seashore and picturesque mountains. The total number of vehicles registered in Ratnagiri is 153491. With the increase in the number of vehicles to 189,619 (Fig. 3) in the year 2013, Ratnagiri has been classified as *slightly polluted* which raises an alarm for quick pollution abatement measures. No city in Maharashtra remains unpolluted as per 2013 pollutant data. The clusters obtained using 2013 air quality data are as follows:

1. Cluster 1 = {} → not polluted
2. Cluster 2 = {Akola, Amravati, Ratnagiri} → slightly polluted
3. Cluster 3 = {Aurangabad, Jalgaon, Kolhapur, Sangli} → moderately polluted
4. Cluster 4 = {Jalna, Latur, Nagpur, Nashik, Solapur} → highly polluted
5. Cluster 5 = {Chandrapur, Mumbai, Pune} → very highly polluted

Table 1 Air quality description of the clusters by inference

Cluster	Air quality description
Not polluted	Very good
Moderately polluted	Good
Poorly polluted	Fair
Very poorly polluted	Poor
Severely polluted	Very poor

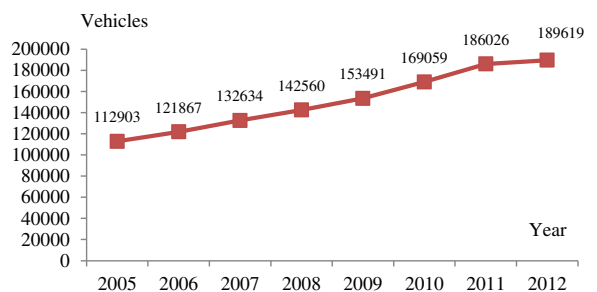


Fig. 3 Vehicle population in Ratnagiri, India

Mumbai and Pune are clustered into a *highly polluted* cluster due to very high concentrations of NO_x and PM₁₀. The reason for increase in PM₁₀ is the alarming increase in number of vehicles in the city which is the main cause of deteriorating air quality. Figure 4 forecasts the pollution load of Ratnagiri city as this city is in the *nonpolluting* cluster but has now been moved to a slightly polluted cluster. The forecast graphs shows that the condition will further worsen if pollution abatement measures are not implemented.

Pollution forecast and application of GA-ZD method

Since majority of cities in India are polluted, it is necessary to visualize future air pollution scenario in a few typical cities. There has been rapid increase in vehicular population resulting into increase in pollution, especially in cities. Pollution forecasted for the selected cities is presented using four methods viz. fuzzy time series, arithmetic increase, incremental increase, and geometric increase.

Pollution load forecast

Figure 5a, b depicts pollution load forecast in Mumbai and New York using four methods of forecasting (http://scetcivil.weebly.com/uploads/5/3/9/5/5395830/m5_15-population_forecasting.pdf). The Chen's method of forecasting makes use of fuzzy time series (Chen 2002; Chen and Hsu 2004). The vehicular pollution forecast for Mumbai city shows that the pollution load is increasing linearly and has already reached an alarming situation and warrants strict pollution abatement measures. New York State has already initiated strict pollution norms in the state for all types

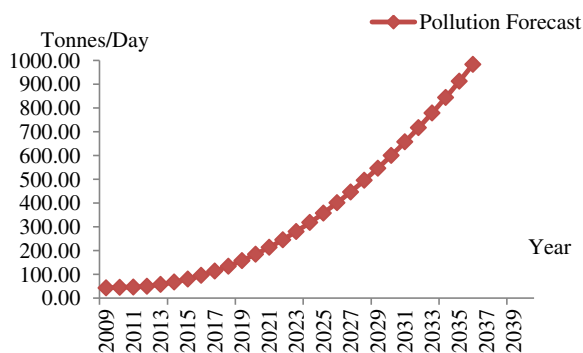


Fig. 4 Ratnagiri pollution load forecast

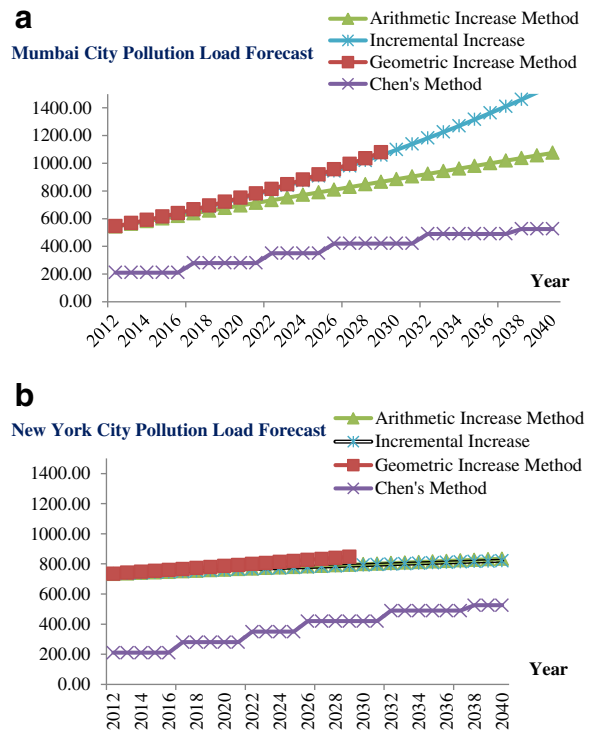


Fig. 5 a Mumbai city pollution load forecast. b New York City pollution load forecast

of vehicles, and thus the graph shows that the pollution is steady and is well under control. The arithmetic mean pollution forecast shows linear increase in pollution, whereas geometric mean shows sudden increase in the pollution load in Mumbai city.

GA-ZD approach

Zadeh–Deshpande approach

Genetic algorithm and Zadeh–Deshpande (GA-ZD) method is the application of fuzzy logic and genetic algorithm in air quality classification. ZD approach directly describes air quality in linguistic terms with linguistic degree of certainty attached to each description. The identification of air quality parameters, number of sampling stations, time and frequency of observations etc., are crucial and are invariably based on the experts' knowledge base. The approach is summarized below:

Collection of pollution parametric data from sampling locations is the first step. On the basis of appropriate probability distribution to the data, mean and variance values could be used in further analysis. When the data is inadequate for statistical distribution fitting, a

well-documented bootstrap method could be employed. Based on the data obtained using bootstrap method, experience has shown that most of the air quality parametric data follows normal distribution. Bootstrap mean and standard deviation are considered while plotting the distribution. As we intend to match probability distribution with the fuzzy sets drawn for the selected parameters (usually possibility distribution), it is necessary to transform probability distribution into possibility distribution for all the criteria pollutants, using the concept of Convex Normalized Fuzzy Number (CNFN).

If x_i is some point on the parametric domain (say in normal distribution) for which $p(x_j)$ is maximum, then define function $\mu_A(x)$ as:

$$\mu_A(x) = p(x_i)/p(x_j) \tag{5}$$

This transforms a random variable into CNFN with membership grade function $\mu_A(x)$, thereby characterizing dynamic behavior of air quality parameters which is a possibility distribution.

Fuzzy sets are generated using GA for selected parameters as per the computational framework explained in “Zadeh–Deshpande (ZD) approach” section computational scheme of degree of match (DM) can be used with a view to estimate matching between fuzzy sets and the antecedent part of the rule, in order to describe air quality fuzzily with certain degree of certainty. A set of fuzzy rules is constructed for classifying air quality as: *very good, good, fair, poor, and very poor* in order to aggregate the set of attributes. The degree of match of each classification rule indicates the certainty value of classification. The greater the degree of match, the greater is the possibility that air quality is classified in that class. The rules are processed using conjunction and disjunction operators. The optimal acceptance strategy is usually that for which the degree of assertion is the maximum (Yadav et al. 2011, 2013, 2014).

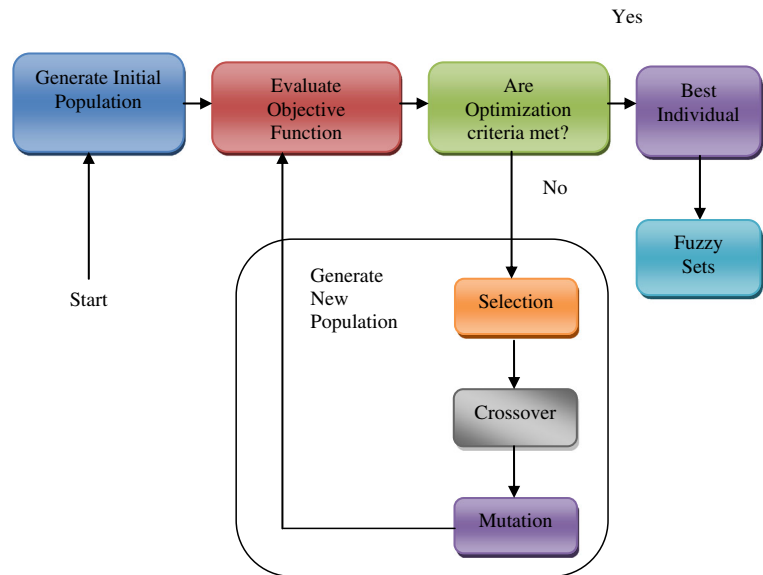
Genetic algorithm

Genetic algorithms were invented by John Holland in 1975 to mimic some of the processes observed in natural selection. The basic concepts of GAs come from Darwin’s theory of evolution, which describes the law of competition and natural selection. By performing the basic operations reproduction, crossover, and mutation, a GA can select better chromosomes from the current population for a specific problem. After the evolution

process, a better next generation can be obtained. Repeating the above process, finally, we can obtain a near-optimal chromosome to deal with a specific problem. In a GA, we encode the parameters of a problem into a numerical string, where the numerical string is called a chromosome. Each element in a chromosome is called a gene. The fitness function is used to evaluate the degree of fitness of a chromosome. The larger the fitness value of a chromosome, the higher the probability of the chromosome to contribute one or more offspring in the next generation. The data type of the genes in a chromosome may be real, integer, or binary values depending on the specific problem (Goldberg 1989; Man et al. 1996; Mitchell 1998).

The initial population is formed by a set of randomly generated members. Each generation consists of members whose constituents are the individual design variables that characterize a design and these are embedded in a binary string. Each member is evaluated using the objective function and is assigned a fitness value, which is an indication of the performance of the member relative to the other members in the population. A biased selection depending on the fitness value decides which members are to be used for producing the next generation. The selected strings are the parents for the next generation, which evolves from the use of two genetic operators namely crossover and mutation.

These operators give a random displacement to the parent population and generate a new population of designs. The crossover operator takes two parent strings, splits them at a random location, and swaps the substrings formed. A probability of crossover determines whether a crossover should be performed. The mutation operator inverts a bit in the string depending on the probability of mutation. The new strings formed are evaluated, and the iteration continues until a maximum number of generations have been reached or until a user-defined termination criterion has been met. Figure 6 shows the sequence of steps in a basic genetic algorithm. The control parameters that have to be initially specified are the population size, the crossover and mutation probabilities, the maximum number of generations, and the termination criterion. Chen and Chen 2010 proposed a new method to generate fuzzy rules from training data by using genetic algorithms. First, the training data is divided into several clusters by using the weighted distance clustering method and generate a fuzzy rule for each cluster. Then, genetic algorithms were used to tune the membership functions of the

Fig. 6 Flowchart for genetic algorithm

generated fuzzy rules. The proposed method attains a higher average classification accuracy rate than the existing methods. Kim et al. 2001 use genetic algorithms to automatically determine the near-optimal rules and their membership functions of fuzzy traffic controllers.

Reproduction (selection) operator The chromosomes with larger fitness values may have more chance to evolve a better next generation. In the reproduction process of a GA, it tries to select the chromosomes that have larger fitness values into a mating pool. The fitness function quantifies the optimality of a solution so that particular solution may be ranked against all other solutions. The function depicts the closeness of the given solution to the desired result. The degree of fitness of each chromosome in a population is computed by using a fitness function. Two commonly used methods of the reproduction operations are roulette wheel and tournament selection operations.

Crossover (recombination) operator Crossover is a genetic operator that combines (mates) two chromosomes (parents) to produce new chromosome (offspring). The idea behind crossover is that the new chromosome may be better than both of the parents if it takes the best characteristics from each of the parents. Crossover selects genes from the parent chromosomes and creates offspring. There are three common methods of crossover operations: one-point crossover, two-point

crossover, uniform crossover operations. The genes of a chromosome are independent and have the same chance to perform the crossover operations.

Mutation operator After crossover is performed, mutation takes place. It is an operator used to maintain genetic diversity from one generation of a population of chromosomes to the next. Mutation occurs during evolution according to a user-definable mutation probability (P_m), usually set to fairly low value, say 0.01 a good first choice. Mutation alters one or more gene values in a chromosome from its initial state. This can result in entirely new gene values been added to the gene pool. With the new gene values, the genetic algorithm may be able to arrive at better solutions than was previously possible. Mutation is an important part of genetic search that helps prevent the population from stagnating at any local optima. Mutation is intended to prevent the search falling into a local optimum of the state space. The flip-bit mutation operator simply inverts the value of the chosen gene (Goldberg 1989; Man et al. 1996; Mitchell 1998).

Computing membership functions using genetic algorithm Genetic algorithms can be used to compute membership functions (Karr and Gentry 1993). Given some functional mapping for a system, some membership functions and their shapes are assumed for the various fuzzy variables defined for a problem. These membership functions are then coded as bit strings that

are then concatenated. An evaluation (fitness) function is used to evaluate the fitness of each set of membership functions (parameters that define the functional mapping).

In a genetic algorithm, the parameter set of the problem is coded as a finite string of bits. The bit strings are combinations of zeros and ones, which represent the value of a number in binary form. An n -bit string can accommodate all integers up to the value $2^n - 1$. For example, the number 7 requires a 3-bit string, i.e., $2^3 - 1 = 7$, and the bit string would look like “111”, where the first unit digit is in the 2^2 place (=4), the second unit digit is in the 2^1 place (=2), and the last unit digit is in the 2^0 place (=1); hence, $4 + 2 + 1 = 7$. The number 10 would look like “1010”, i.e., $2^3 + 2^1 = 10$, from a 4-bit string. This bit string may be mapped to the value of a parameter, say C_i , $i = 1, 2$, by the mapping:

$$C_i = C_{min} + \frac{b}{2^{L-1}}(C_{max_i} - C_{min_i}) \tag{6}$$

where b is the number in decimal form that is being represented in binary form (e.g., 152 may be represented in binary form as 10011000), L is the length of the bit string (i.e., the number of bits in each string), and C_{max} and C_{min} are user-defined constants between which C_1 and C_2 vary linearly. The parameters C_1 and C_2 depend on the problem.

First, an initial population of n strings (for n parameters) of length L is created. The strings are created in a random fashion, i.e., the values of the parameters that are coded in the strings are random values (created by randomly placing the zeros and ones in the strings). Each of the strings is decoded into a set of parameters that it represents. This set of parameters is passed through a numerical model of the problem space. The numerical model gives out a solution based on the input set of parameters. On the basis of the quality of this solution, the string is assigned a fitness value. The fitness values are determined for each string in the entire population of strings. With these fitness values, the three genetic operators are used to create a new generation of strings, which is expected to perform better than the previous generations (better fitness values). The new set of strings is again decoded and evaluated, and a new generation is created using the three basic genetic operators. This process is continued until convergence is achieved within a population.

Among the three genetic operators, reproduction is the process by which strings with better fitness values

receive correspondingly *better copies* in the new generation, i.e., we try to ensure that better solutions persist and contribute to better offspring (new strings) during successive generations. This is a way of ensuring the “survival of the fittest” strings. Because the total number of strings in each generation is kept a constant (for computational economy and efficiency), strings with lower fitness values are eliminated.

The second operator, crossover, is the process in which the strings are able to mix and match their desirable qualities in a random fashion. After reproduction, crossover proceeds in three simple steps. First, two new strings are selected at random. Second, a random location in both strings is selected. Third, the portions of the strings to the right of the randomly selected location in the two strings are exchanged. In this way, information is exchanged between strings, and portions of high-quality solutions are exchanged and combined. Reproduction and crossover together give genetic algorithms most of their searching power. The third genetic operator, mutation, helps to increase the searching power (Ross 2009).

Describing air quality The case study relates to describing air quality in linguistic terms using parametric data for Mumbai city in India and New York City in the USA (Fig. 7a, b). Maharashtra Pollution Control Board (MPCB) monitors three pollutants viz. particulate matter (PM_{10}), oxides of nitrogen (NO_x), and oxides of sulfur (SO_x) (<http://www.mpcb.gov.in/> Retrieved on 1/10/2013) which were considered for linguistic description of air quality at Mumbai in two locations viz. Sion and Bandra. New York city monitors carbon monoxide (CO), particulate matter ($PM_{2.5}$), and ozone (O_3) and were thus considered to describe air quality at New York (<http://www.dec.ny.gov/airmon/> Retrieved on 10/10/2013) in four locations viz. City College New York (CCNY), Division Street, Public School (PS)-19, and International School (IS)-143. The parametric data was considered for the winter month of November 2013 for both the cities. A computational scheme of degree of match (DM) is used with a view to estimate matching between the assertion and the antecedent part of fuzzy rules, in order to describe air quality fuzzily with linguistic degree of certainty.

Reliability of monthly mother sample data of the pollutants and the identification of the worst winter month has been ensured using bootstrapping. One thousand bootstrap samples were compared for mean and

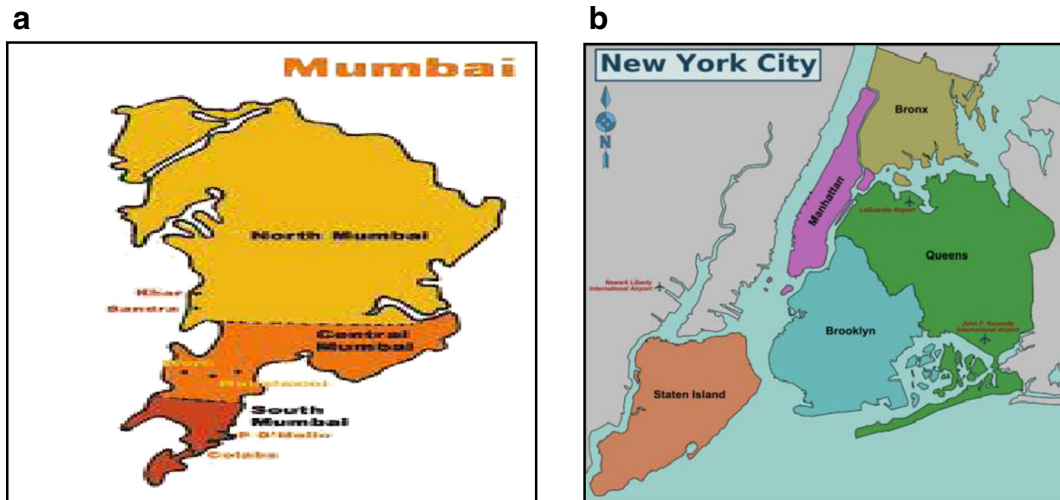


Fig. 7 a Mumbai (India). b New York (USA)

standard error of the original and the bootstrap samples. In the present study, bootstrap mean for the month of November 2013 for all the pollutants monitored in Mumbai and New York is considered for further analysis. CNFN was constructed for the data sets of the parameters and linguistic description of air quality was obtained from the experts. GA was applied to construct fuzzy sets describing the interval of confidence for the terms very good, good, fair, poor, and very poor.

Tables 2 and 3 portray the DM between the parametric data of the pollutants computed at monitoring stations in New York and Mumbai cities and GA generated fuzzy sets in linguistic hedges for these parameters. For example, at Sion station, the highest DM for SO_x, NO_x,

and PM₁₀ are 1, 1, and 0.96 for the linguistic description *very good*, *very poor*, and *very poor*, respectively, and for Bandra station, the highest DM is 1 for all SO_x, NO_x, and PM₁₀ for the linguistic description *very good*, *poor*, and *very poor*, respectively. In case of New York, the highest DM at station CCNY is 1 for all pollutants O₃, PM_{2.5}, and CO for linguistic description *very good*.

Based on the parametric data and fuzzy sets generated by GA (Table 4), it can be concluded that the air quality in Mumbai is generally described as *very poor*. Table 4 describes air quality linguistically with linguistic degree of certainty using GA generated fuzzy sets. It can be concluded that the AQ at all the monitoring stations in New York is *very good* with degree of certainty as *very high*. The air quality at Sion city in Mumbai can be directly described linguistically as *very poor* with a *very high* (0.95) degree of certainty value. The next higher

Table 2 Degree of match—New York

Degree of match (New York)					
City College New York (CCNY)					
Pollutants	VG	G	F	P	VP
O ₃	1	0	0	0	0
PM _{2.5}	1	0	0	0	0
CO	1	0	0	0	0
Division Street					
PM _{2.5}	1	0	0	0	0
Public School (PS)-19					
PM _{2.5}	1	0	0	0	0
International School (IS)-143					
PM _{2.5}	1	0	0	0	0

Table 3 Degree of match—Mumbai

Degree of match (Mumbai)					
Sion					
Pollutants	VG	G	F	P	VP
SO _x	1	0	0	0	0
NO _x	0	0	0.5	0.88	1
PM ₁₀	0	0	0.23	0.46	0.96
Bandra					
SO _x	1	0	0	0	0
NO _x	0	0	0.52	1	0.76
PM ₁₀	0	0	0	0.82	1

Table 4 Linguistic description of air quality and degree of certainty

Linguistic class	New York City				Mumbai city	
	CCNY	Division street	PS-19	IS-143	Sion	Bandra
AQ description with GA generated fuzzy sets						
VG	1 DC-VH	1 DC-VH	1 DC-VH	1 DC-VH	0	0
G	0	0	0	0	0.29	0
F	0	0	0	0	0.46	0
P	0	0	0	0	0.78	0.23
VP	0	0	0	0	0.95 DC-VH	0.93 DC-VH
AQ description with expert fuzzy sets (perception based)						
VG	1 DC-VH	1 DC-VH	1 DC-VH	1 DC-VH	0	0
G	0	0	0	0	0	0
F	0	0	0	0	0.23	0
P	0	0	0	0	0.46	0.82
VP	0	0	0	0	0.96 DC-VH	1 DC-VH

value of 0.78 for *poor* indicates that the air quality at Sion is definitely deteriorating as the DC is 0.78 which is higher than 0.46 for the linguistic hedge *fair*. In the present scenario, the air quality description of both cities using expert fuzzy and GA generated fuzzy sets is same. There is no change in the final description of air quality. Thus, genetic algorithm could be a better choice only in case where the domain experts are unavailable as far as air quality is concerned.

The lower part of Table 4 describes air quality linguistically with linguistic degree of certainty using fuzzy sets based on the perceptions of the domain experts. Using expert fuzzy sets, the air quality at New York could be linguistically described as *very good* with degree of certainty as *very high* (1). The air quality at Sion city in Mumbai can be directly described linguistically as *very poor* with a *very high* (0.96) degree of certainty value. The next higher value of 0.46 for *poor* indicates that the air quality at Sion is definitely deteriorating as the DC is 0.46 which is higher than 0.23 for the linguistic hedge *fair*. It is pertinent to mention that fuzzy logic approach attaches certainty values to the linguistic terms while describing air quality, whereas the conventional method describes air quality index first with a numeric value and then describes it in linguistic terms which is a departure from human thinking.

High concentration of PM₁₀ in air can be attributed to heavy vehicular traffic, construction activities, cement roads, untarred roads, etc. in Mumbai city. Low

temperature and less vertical dispersion during winter months increase the concentration of PM₁₀. The fact that there is not much horizontal air movement in summer also leads to localization of PM₁₀. During summer, there is lack of convection winds and particulate matter gets concentrated in the area where they are generated and are not dispersed. The increase in the number of vehicles has eventually led to alarming levels of emissions as they are caught at the intersections and in the usual snarls during peak rush hours. The flyovers constructed for fast movement vehicles only added auto exhaust emissions. Industrial units located in cities add to oxides of nitrogen, SO_x, and particulate matter. The authors believe that the situation will further deteriorate if the present trend of plying more and more polluting vehicle and construction activities in Mumbai continues. From Table 5, it can be inferred that the pollution in New York City is within limits. The average values of NO₂ and PM₁₀ for Sion are 177.78 and 151.26 µg/m³, respectively, which are higher than the permissible values of 80 and 100 µg/m³ as per Table 6 for 24 h average. Tables 6 and 7 present the NAAQS for Mumbai and New York, respectively. In New York, the air quality can be described as *very good* with *very good degree of certainty*.

Recognizing that unintentional carbon monoxide poisoning is a serious but preventable environmental health threat, New York City enacted a law in 2004 requiring CO alarms in residential and many public buildings and updated the Health Code to make CO

Table 5 New York and Mumbai: air quality criteria pollutants: basic statistics

New York CCNY	O ₃ (µg/m ³)	PM _{2.5} (µg/m ³)	CO (µg/m ³)
Minimum	9.82	3.6	56.46
Maximum	34.36	16.36	792.84
Average	21.03	7.02	351.96
Std Dev (σ)	6.8	2.9	219.21
Avg±3σ	0.61 to 41.44	-1.69 to 15.72	-305.67 to 1009.6
Mumbai SION	SO ₂ (µg/m ³)	NO _x (µg/m ³)	PM ₁₀ (µg/m ³)
Minimum	4	114	69
Maximum	5	268	242
Average	4.09	177.78	151.26
Std Dev (σ)	0.29	39.5	44.16
Avg±3σ	3.22 to 4.95	59.3 to 296.27	18.78 to 283.74

poisoning immediately reportable by telephone to the Department of Health and Mental Hygiene/NYC Poison Control Center.

New York City has initiated strict pollution control norms and enacted a law in 2004 requiring CO alarms in residential and many public buildings. New York City's air quality has improved over time as the regulations have made federal, state, and local air quality standards more stringent over the last two decades. Federal and state regulatory efforts to reduce emissions from the

transportation, off-road, and stationary source sectors have driven continued national improvements in air quality. As required by the federal Clean Air Act, the US Environmental Protection Agency (US-EPA) sets standards for particulate matter (PM_{2.5}), nitrogen dioxide (NO₂), and sulfur dioxide (SO₂) emissions from

Table 6 Air quality standards established by CPCB

Pollutant (µg/m ³)	Time-weighted average	Industrial, residential, rural, and other area
Sulfur dioxide (SO ₂)	Annual	50
	24 h	80
Nitrogen dioxide (NO ₂)	Annual	40
	24 h	80
Particulate matter PM ₁₀	Annual	60
	24 h	100
Particulate matter PM _{2.5}	Annual	40
	24 h	60
Ozone (O ₃)	8 h	100
	1 h	180
Lead (Pb)	Annual	0.50
	24 h	1.0
Carbon monoxide (CO)	8 h	02
	1 h	04
Ammonia (NH ₃)	Annual	100
	24 h	400

Source: http://www.cpcb.nic.in/National_Ambient_Air_Quality_Standards.php

Table 7 Air quality standards established by US-EPA

National ambient US-EPA air quality standards		
Pollutant	Standard value	
Carbon monoxide (CO)		
8-h average	9 ppm	10,000 µg/m ³
1-h average	35 ppm	40,000 µg/m ³
Nitrogen dioxide (NO ₂)		
Annual arithmetic mean	0.053 ppm	100 µg/m ³
Ozone (O ₃)		
1-h average	0.12 ppm	235 µg/m ³
8-h average	0.08 ppm	157 µg/m ³
Lead (Pb)		
Quarterly average	–	1.5 µg/m ³
Particulate (PM ₁₀)		
Annual arithmetic mean	–	50 µg/m ³
24-h average	–	150 µg/m ³
Particulate (PM _{2.5})		
Annual arithmetic mean	–	15 µg/m ³
24-h average	–	65 µg/m ³
Sulfur dioxide (SO ₂)		
Annual arithmetic mean	0.030 ppm	(80 µg/m ³)
24-h average	0.14 ppm	(365 µg/m ³)
3-h average	0.50 ppm	(1300 µg/m ³)

Source: <http://epa.gov/air/criteria.html>

large fossil-fuel combustion sources and also designates areas that fail to meet health-based standards for air quality established by the agency. Most recently, between 2006 and 2010, the US-EPA phased in stringent emissions standards for heavy duty diesel vehicles such as trucks and for gasoline passenger vehicles while also reducing the sulfur content of diesel fuel and gasoline. These new emissions standards, once fully implemented, are expected to significantly reduce air pollution from key transportation sectors and reduce their associated public health impacts, thus contributing to improved air quality in New York City.

From the air quality status estimated by GA-ZD approach combined, it can be concluded that the air quality in Mumbai is *very poor* with *very high* degree of certainty and that of New York, it is *very good* with *very high* degree of certainty. The air pollution authorities in Mumbai, India, have to follow stringent pollution control norms as the vehicle population is growing at an alarming rate in Mumbai. Petrol/diesel-driven vehicles should be discouraged. The policy on plying nonpolluting vehicles should be implemented.

Software support

A system for linguistic description of air quality, including web-based air quality description software has been developed. The website takes as input the concentrations of three pollutants viz. NO_x , SO_x , and PM_{10} in $\mu\text{g}/\text{m}^3$ and displays as output a facial expression to describe the status of air. The software is developed using the fuzzy-logic-based formalism in ZD approach that models aleatory and epistemic uncertainty which is not taken care of in the conventional air quality index method. The air quality computations are performed using a web-based software developed on Windows Operating System platform. The web server used is Apache or Nginx. MySQL (structured query language) is used as a relational database management system and for the client-server architecture. The database server (MySQL) and the client (application program) communicate with each other for querying data, saving changes, etc. The technology and tools used for development are Hypertext Preprocessor (PHP5) which is a server-side scripting language designed for web development or for a general purpose programming

language, HyperText Markup Language (HTML) is the main markup language used for creating WebPages and information to be displayed on the web browser, Cascading Style Sheets (CSS) is a style sheet language used for describing presentation semantics of a document designed in HTML, Java Query (JQuery) is a multi-browser JavaScript library designed to simplify the client-side scripting of HTML, and JavaScript (JS) is an interpreted computer programming language. It is implemented as a part of web browser so that client-side scripts can interact with the user, control the browser, communicate asynchronously, and alter the document content. The code contains 22 PHP files with approximately 1400 lines of code, 75 JS files with 3500 lines of code, and 822 CSS files with 1000 lines of code. The computation time of approximately 2 min is required only for data entry which includes inputting three pollutant values. The final output which describes air quality linguistically with linguistic degree of certainty is obtained on the click of the buttons. The site address for the software is <http://www.jyoti Yadav.in/>. However, no tall claims are made and the website can be modified to include more number of air pollutants and expert fuzzy sets for those pollutants.

Conclusion and future scope for research

The case studies presented infers the strength of fuzzy set theoretic approach in combination with genetic algorithm in straightway describing air quality in linguistic terms with linguistic degree of certainty attached to each term. A close look at the limited study will show that the fuzzy sets drawn on the basis of expert knowledge base are not much different than those obtained using genetic algorithm.

The classification of air quality monitoring stations using Fuzzy C mean clustering algorithm in association with Zadeh–Deshpande could be a better method in decision making for initiating pollution abatement action plans. The authors have demonstrated the concept of reference group and computed weight factor value which is needed in Fuzzy C mean clustering. Having said this, there is a future need to use other soft computing techniques in air quality modeling pollution assessment, pollution risk,

and for suggesting pollution mitigation measures. There is a long way to go.

References

- Bezdek, J., Ehrlich, R., & Full, W. (1984). FCM: the fuzzy c-means clustering algorithm. *Computers and Geosciences*, 10(2–3), 191–203.
- Brown, E., Nichols, M., Corey, W. (2006). *First update to the climate change scoping plan building on the framework*. The California Global Warming Solutions Act.
- Chen, S. (2002). Forecasting enrolments based on high-order fuzzy time series. *Cybernetics and Systems: An International Journal*, 33, 1–16.
- Chen, S. M., & Chen, Y. C. (2010). Automatically constructing membership functions and generating rules using genetic algorithm. *Cybernetics and Systems: An International Journal*, 33(8), 841–862.
- Chen, S., & Hsu, C. (2004). A new method to forecast enrolments using fuzzy time series. *International Journal of Applied Science and Engineering*, 2(3), 234–244.
- Goldberg, D. (1989). *Genetic algorithms in search, optimization, and machine learning, Chapter 1–8* (pp. 1–432). Reading: Addison Wesley.
- Karr, C. L., & Gentry, E. J. (1993). Fuzzy control of pH using genetic algorithms. *IEEE Transactions on Fuzzy Systems*, 1(1), 46–53.
- Kim, J., Kim, B.M., Kim, NCH. (2001). Genetic algorithm approach to generate rules and membership functions of fuzzy traffic controller. *IEEE Xplore*. O-7803-7293-X/01/\$17.00, 525-528.
- Man, K., Tang, K., & Kwong, S. (1996). Genetic algorithm: concepts and applications. *IEEE Transactions in Industrial Electronics*, 43(5), 519–534.
- Mitchell, M. (1998). *An introduction to genetic algorithm. Chapter 1–6* (pp. 1–203). Cambridge: MIT Press.
- Ross, T. (2009). *Fuzzy logic with engineering applications* (2nd ed., pp. 369–375). India: Wiley.
- Saksena, S., Joshi, V., & Patil, R. (2003). Cluster analysis of Delhi's ambient air quality data. *Journal of Environmental Monitoring*, 5, 491–499.
- Yadav, Y., Kharat, V., Deshpande A. (2011). *Fuzzy description of air quality: a case study*. 6th International conference on Rough Sets and Knowledge Technology (RSKT), Banff, Canada, Oct. 9–12, 420–427.
- Yadav, J., Kharat, V., & Deshpande, A. (2013). Evidence theory and fuzzy relational calculus in estimation of health effects due to air pollution. *International Journal of Intelligent Systems*, 22(1), 9–22.
- Yadav, J., Kharat, V., Deshpande, A. (2014). Fuzzy description of air quality using fuzzy inference system with degree of match via computing with words: a case study. *International Journal of Air Quality, Atmosphere and Health*.