Chapter 5

Soft Computing Techniques for Air Quality Classification

Prologue

With a focus on health risk, one must keep in mind that what concerns us is not the level of pollution itself, but rather the level that reaches the lungs—a distinction that fundamentally motivates and shapes this assessment. In this regard, Chapter 4 presented a method based on evidence theory and fuzzy relational calculus to associate polluted air with respiratory diseases. The method is not epidemiologically data intensive and models only the perceptions of the Chest Physicians/Pulmonologists for associating polluted air with respiratory diseases. In the current chapter detailed description of air quality assessment studies, carried out primarily using Zadeh-Deshpande fuzzy logic based formalism, has been presented.

From the human health risk viewpoint it becomes necessary to measure, first, the concentrations of criteria air pollutants in the atmosphere to describe the status of air that we breathe. If the concentration exceeds the prescribed limits, then implementation of air pollution abatement measures is the most important task for the pollution control agencies across the globe.

One of the ways to classify air is with indices. Air Quality Index (AQI) is a numeric value which is classified or described linguistically as poor, fair, good air quality and conveyed to the public at large. As the AQI increases, it is hypothesized
that a sizeable percentage of the population is likely to experience increasingly severe adverse health effects. The method suggested by US-EPA for computation of air quality indices is effective. However, the method suffers from a serious drawback as it does not include domain expert’s knowledge which is invariably based on partial belief. In India, all the pollutants are not measured under the National Ambient Air Quality Monitoring Programme (NAAMP). The regulatory authorities (Central Pollution Control Board) suggested measuring at least three criteria pollutant concentrations for collating air quality index.

The overall scheme of the research is centred on application of Fuzzy-Genetic (Fuzzy-GA) techniques to linguistically describe air quality with a degree of certainty attached to each linguistic description. It is important to mention herein that the potential of genetic algorithm is explored only to the extent of the construction of fuzzy sets in this research.

The following section presents the salient features of various case studies reported in this document. Comparative evaluation of the results obtained using the soft computing-based methods are made with the conventional method of computing air quality index and is presented in Chapter 6.

### 5.1 Fuzzy Description of Air Quality and Perception-Based Modelling

This section presents salient features of the case studies carried out by applying the ZD formalism which directly describes air quality linguistically. More than one expert’s perceptions were considered for the analysis wherein the variability in their perception was evident from the fuzzy sets (see Annexure-1) and fuzzy rule base (see Annexure-2). The variability in the experts’ perception has been modelled using fuzzy logic.
5.1.1 Case Study 1

The case studies relates to monitoring locations in Maharashtra and Tamil Nadu states from India. The Maharashtra Pollution Control Board (MPCB) and Tamil Nadu Pollution Control Board (TNPCB) monitor three pollutants viz. Particulate Matter (PM$_{10}$), Oxides of Nitrogen (NO$_X$) and Oxides of Sulphur (SO$_X$). Monitoring of these three pollutants shows that these pollution control boards attach more importance to the pollution due to vehicular traffic. It is true that the increase in PM$_{10}$ could also be attributed to ongoing construction activities.

The research is an attempt to compare the following three methods for air quality classification using the available air quality parametric data in four Indian cities viz. Mumbai, Navi Mumbai, Pune and Chennai.

The selective methods in air quality classification include:

1. Conventional Air Quality Index (CAQI) formulated by US-EPA.
2. Fuzzy Air Quality Index (FAQI) using Fuzzy Inference System.
3. Zadeh-Deshpande (ZD) Formalism

The first two methods fail to capture the uncertainty which is resident in the pollutant data. In these two methods, first the air quality is computed as a number and then described in linguistic terms, unlike human thinking. The study reveals that the air quality description using the third formalism is comparable to CAQI and FAQI methods. In addition, the advantage of the third method is that it attaches a degree of certainty to each linguistic class. Figure 5.1 depicts some of the facets of the three methods.

The first part of the case study relates to fuzzy description of air quality with the available air quality data from five monitoring stations in Pune city viz. Karve Road, Swargate, Nal Stop, Bhosari and PCMC, three monitoring stations in Mumbai...
Chapter 5: Case Study 1

viz. Sion, Mulund and Bandra and four monitoring stations in Navi Mumbai viz. Vashi, Airoli, Nerul and Mahape.

Figure 5.1 Framework for Computing AQI

Data from five locations for five years, from 2008 to 2012 and for six months between Jan–Mar and Oct–Dec, was considered to model the variability in the experts’ perceptions. The second part of the case study relates to describing air quality using parametric data from five air quality monitoring stations in Chennai city of Tamil Nadu state viz. Anna Nagar, Adyar, Kilpauk, Theagaraya Nagar (T. Nagar) and Vallalar Nagar (V. Nagar). The air quality data for the winter months of 2013 was obtained from TNPCB website and was considered to model the uncertainty in the
experts’ perception. The parametric data for the worst months in winter has been considered in both the cases when temperature inversion is observed.

A computational scheme of Degree of Match (DM) is used with a view to estimate the match between the assertion and the antecedent part of fuzzy rules, in order to describe air quality fuzzily with a degree of certainty.

The bootstrap is a **re-sampling method** for assessing uncertainty. It is commonly used to estimate confidence intervals, but it can also be used to estimate bias and variance of an estimator or calibrate hypothesis tests. Bootstrapping was applied on the parametric data which was then converted to probability density function. For example, consider the observed data for PM$_{10}$ obtained from the TNPCB website as 139,138,129,121,110,66,52,53,46,37,62,49,48,45,50,63,73,94,84,121,129,121,120,130,133 which had 25 values. Bootstrapping was performed on these sets of values, on 1000 re-sample values. Average ($\mu$) and standard deviation ($\sigma$) were computed. It was observed that there was no significant difference in the sample standard error and bootstrap standard error. Frequency graph was plotted which showed that the bootstrapped data somewhat follows normal distribution. Rigorous significant tests were not carried out for proving that the data follows normal distribution. A probability density function with values ($\mu - 3\sigma, \mu, \mu + 3\sigma$) was plotted. As we need to compare probability density function drawn with the expert fuzzy sets we convert the PDF into possibility distribution. Figure 5.2 depicts the probability and possibility distribution plot along with the Convex Normalized Fuzzy Number (CNFN) for the PM$_{10}$ parameter. Bootstrapping computations are included in the **Annexure-3**.

Figure 5.3 presents the degree of Match of the field data for PM$_{10}$ with linguistic term ‘Poor’ at Pimpri-Chinchwad Municipal Corporation (PCMC) monitoring station. The phenomenal increase in the number of two wheelers has caused a serious problem of vehicular air pollution in cities like Pune and Mumbai. The total pollution load in Pune city due to 1,784,740 vehicles is 479 tonnes per day.
Chapter 5: Case Study 1

The vehicular pollution due to 2 wheelers is the maximum, which is 30.28 % of the total pollution load.

![Figure 5.2 Probability and Possibility Distribution for Particulate Matter (PM$_{10}$)](image1)

Figure 5.2 Probability and Possibility Distribution for Particulate Matter (PM$_{10}$)

![Figure 5.3 Degree of Match of field data on PM$_{10}$ with Linguistic term Poor (PCMC)](image2)

Figure 5.3 Degree of Match of field data on PM$_{10}$ with Linguistic term Poor (PCMC)
Chapter 5: Fuzzy Description of Air Quality using FIS with DM via CW and Classification of Polluted Cities using FCM

Total Pollution load in Mumbai city due to 1,784,740 vehicles is 665 tonnes per day. Pollution due to cars is the maximum which is 28.94% of the total pollution load. The two wheelers are the main cause of vehicle pollution in Pune city and diesel driven four wheelers (cars) are the main cause of pollution in Mumbai city. The auto exhaust pollutants viz. PM$_{10}$ and NO$_X$ are of great concern to the environmentalists.

5.2 Fuzzy Description of Air Quality using FIS with DM via CW and Classification of Polluted Cities using FCM

In addition to fuzzy description of air quality, in this section the salient features of the data needed in computation of Fuzzy C Means to classify 14 cities in Maharashtra into five different clusters based on the air pollution level are presented.

5.2.2 Case Study 2

The first part of the case study examines the status of ambient air quality in 14 cities (Figure 5.4) with 51 monitoring locations from Maharashtra, India. Parametric data of November 2011 has been considered to convey air quality status to public.
Chapter 5: Case Study 2

Air quality has been computed using CAQI, FAQI and ZD approach, this time laying focus on Computing with words. The ZD method describes air quality straightway in a linguistic term with *linguistic* degree of certainty. Uncertainty of uncertainty is modelled using type I fuzzy logic. Firstly, the uncertainty in the expert’s perception is modelled and then the second uncertainty related to a degree of certainty is modelled by defining fuzzy sets for Degree of Certainty as described in the graph (Figure 5.5) below. The graph depicts five fuzzy sets as *Very Good, Good, Fair, Poor and Very Poor* for linguistic classification of Degree of Certainty. The inherent strength of the method is linguistic description of air quality with linguistic degree of certainty.

The second part of the case study relates to clustering of fifteen cities of Maharashtra (India), based on their pollution potential, into five different clusters. Classification of fifteen cities each described with two features viz. NO\textsubscript{X} and PM\textsubscript{10} has been considered. The fifteen cities 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)*.

Figure 5.5 Fuzzy Sets for Degree of Certainty
5.3 Pollution Forecast and GA-Fuzzy Modelling in Air Quality Classification

This section focuses on the application of Genetic Algorithm to generate fuzzy sets for various air pollutants and finding their similarity with the fuzzy sets defined by the air quality experts’.

The concept of similarity is fundamentally important in almost every scientific field which helps us to know how much two things are similar. For this we, first calculate the degree of similarity. Higher the degree of similarity between two things, they are more similar to each other. The degree to which people perceive two things as similar fundamentally affects their rational thought and behavior. There are a large number of similarity measures proposed in the literature. All similarity measures should map to the range [-1, 1] or [0, 1]. 0 and -1 shows minimum similarity and 1 shows maximum similarity. There are three models of similarity measures: Distance based, Feature based and Probabilistic similarity measures (Beg 2009).

Let us assume that the universe of discourse U to be finite set and \( A = \sum_{u \in U} \mu_A(u) \) and \( B = \sum_{u \in U} \mu_B(u) \) be two fuzzy sets defined over U. A similarity index between the pair (A, B) is denoted as \( S(A, B) \). In order to compare the fuzzy sets defined by the air quality experts’ (Figure 5.6) and those generated using Genetic algorithm (Figure 5.7) following 5 similarity measures have been applied (Gupta 2014).

1. \( S(A, B) = \sqrt{\sum_{i=1}^{n} (\mu_{A_i} - \mu_{B_i})^2} \)

2. \( (A, B) = 1 - \left[ \frac{\sum_{u \in U} |\mu_{A(u)} - \mu_{B(u)}|^q}{n^q} \right]^{\frac{1}{q}} \); q: family parameter, n: Universe of discourse

3. \( S(A, B) = 1 - \max_{u \in U} (|\mu_A(u) - \mu_B(u)|) \)
Chapter 5: Pollution Forecast and GA-Fuzzy Modelling in Air Quality Classification

4. \[ S(A, B) = \frac{1}{1 + \sqrt{\sum_{i=1}^{n} (\mu_{A_i} - \mu_{B_i})^2}} \]

5. \[ S(A, B) = \frac{\sum_{u \in U} |\mu_{A}(u) - \mu_{B}(u)|}{\sum_{u \in U} |\mu_{A}(u) + \mu_{B}(u)|} \]

Figure 5.6 Expert Fuzzy set for PM\textsubscript{10} Linguistic Variable Very Good

Figure 5.7 GA Generated Fuzzy set for PM\textsubscript{10} Linguistic Variable Very Good
A trapezoidal membership function is specified by four parameters \( \{a, b, c, d\} \) as follows:

\[
\text{Trapezoid}(X:a,b,c,d) = \begin{cases} 
0 & ; x < a \\
\frac{x - a}{b - a} & ; a \leq x < b \\
\frac{1}{b - a} & ; b \leq x \leq c \\
\frac{d - x}{d - c} & ; c < x \leq d \\
0 & ; x > d
\end{cases}
\]

Consider the Expert fuzzy set for \textit{Very Good} as given in the graph of Figure 5.6

\[
A = \left\{ \frac{0}{18} + \frac{1}{25} + \frac{1}{42} + \frac{0}{57} \right\}
\]

And GA generated Fuzzy Set for \textit{Very Good} as given in the graph of Figure 5.7

\[
B = \left\{ \frac{0}{18} + \frac{0.5}{25} + \frac{1}{42} + \frac{0.2}{57} \right\}
\]

Using the fuzzy sets A and B the degree of similarity is computed in Table 5.1 using the above 5 similarity measures.

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Fuzzy Set for PM_{10}</th>
<th>Similarity Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Method 1</td>
</tr>
<tr>
<td>1</td>
<td>Very Good</td>
<td>0.54</td>
</tr>
<tr>
<td>2</td>
<td>Good</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>Fair</td>
<td>0.62</td>
</tr>
<tr>
<td>4</td>
<td>Poor</td>
<td>0.55</td>
</tr>
<tr>
<td>5</td>
<td>Very Poor</td>
<td>0.65</td>
</tr>
</tbody>
</table>

The degree of similarity values in the above table conclude that the two fuzzy sets viz. Expert fuzzy sets and the fuzzy sets generated by using Genetic Algorithm
are similar with degree of similarity ranging from 0.50 to 1.00. Method 5 gives the highest degree of similarity. The Very Good fuzzy sets of the experts and the one generated by GA have the degree of similarity as 0.92. From these values we can also conclude that the final outcome of linguistically describing air quality using ZD formalism is almost similar using expert fuzzy sets and GA generated fuzzy sets. The authors still believe in the perception of the experts rather than generating fuzzy sets using Genetic Algorithm wherein the final result is not affected.

Since majority of the cities in India are polluted, it is necessary to visualize future air pollution scenario in a few typical cities. There has been a rapid increase in vehicular pollution resulting to 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. Figure 5.8 gives the approach used for linguistically describing air quality forecasting pollution and clustering polluted cities into different clusters.

5.3.1 Case Study 3

Two metropolitan cities viz. Mumbai in India and New York in USA are identified for the assessment of their pollution status due to their somewhat similar geographical features. As the first part of the case study, Figure 5.9 and 5.10, depict forecast of pollution load in Mumbai and New York using four methods of forecasting. The Chen’s method of forecasting makes use of fuzzy time series. The vehicular pollution forecast for Mumbai city (Figure 5.9) shows that the pollution load is increasing linearly and has already reached an alarming situation and warrants strict pollution abatement measures. The arithmetic mean pollution forecast shows linear increase in pollution whereas geometric mean shows sudden increase in the pollution load in Mumbai city. New York State has already initiated strict pollution norms in the state for all types of vehicles and thus the graph shows that the pollution is steady and is well under control (Figure 5.10).
The second part of the case study is GA-Fuzzy modelling, in combination with ZD formalism, is applied to assess and compare the air quality status of New York and Mumbai (Figure 5.11 and 5.12). 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. The parametric data for the winter month of November 2013 has been 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 in four locations viz. City College New York (CCNY), Division Street, Public School(PS)-19, and International School(IS)-143. The same computational procedure as described
in case study 1 has been applied, except for the fuzzy sets. The fuzzy sets generated by applying genetic algorithm have been used. GA was applied to construct fuzzy sets describing the interval of confidence for the terms Very Good, Good, Fair, Poor and Very Poor.

Figure 5.11 Mumbai (India)  
Figure 5.12 New York (USA)

Chapter 5 thus discusses the case studies carried out in the classification of air quality by using perception-based modelling. Chapter 6 refers to a detailed discussion of results obtained in the defined studies. A closer look at the results of the case studies will indeed highlight the importance of fuzzy logic in describing air quality straightaway in linguistic terms with degree of certainty attached to each description. This is somewhat like Z-Number proposed by Professor Lotfi Zadeh in his relatively new concept of computing with words.