Chapter 4

Research Methodology

This chapter covers the research methodology, adopted for the study. The problem definition and gap areas are identified in previous chapters have provided direction and guidance for the research objectives.

4.1 Research Methodology

The focus of this research is on measuring and predicting the level of air pollutants and their trends in Mumbai, Navi Mumbai and nearby areas. The research work consists of three parts. In the first part, the linkage between the level of air pollution and the various meteorological parameters has been studied. In the second part, the levels and trends of air pollution in Mumbai and Navi Mumbai city has been analyzed and modeled. In the third part, kriging models are used to predict pollution levels in nearby areas.

The overall steps in the research are as follows. The spatial data and attribute data was collected from various sources. The data was edited, analyzed and formatted as per the requirements of the data mining and kriging. The data mining and kriging tools developed were used to generate the desired predictions and kriging outputs. Finally, the results were displayed and printed as per user requirements. The details of steps, tools and techniques used are now explained. The methodology followed to answer the above questions is as follows:
1. Collection of spatial and attribute data from MET and MPCB department
2. Explore data.
3. Analyze data using Data Mining Techniques
4. Use the supervised learning to mine the hidden knowledge embedded in the database.
5. Display the mining result manually or automatically explain why they happen.
6. Mined data from MET database is the source for predicting the pollutant concentration.
7. The prediction is done using ANN computational model, which brings the link between MET parameters and selected pollutants.
8. For the areas where MET stations are unavailable, Kriging interpolation technique is used to interpolate pollutant concentration.

4.2 Data Collection

The research work is primarily based on data collected from the following authentic sources:

1. Maharashtra Pollution Control Board: MPCB is an autonomous statutory body. In certain administrative and technical matters, it is responsible to State Government, Central Pollution Control Board and Ministry of Environment and Forests, Govt. of India.

2. Meteorological department is an autonomous statutory body under the State Government. It is responsible for collecting and maintaining meteorological data at station level and sea level.

3. National Environmental Engineering Research Institute (NEERI), Nagpur is a constituent of Council of Scientific and Industrial Research (CSIR). It has five zonal laboratories at Chennai, Delhi, Hyderabad, Kolkata and Mumbai.
4. Municipal Corporation of Greater Mumbai (MCGM) collects data for parameters viz. SO$_2$, NO$_2$, RSPM, at sites viz. traffic junctions, dumping sites, industrial, residential societies etc.


The data for air pollutants (SO$_x$, NO$_x$ and RSPM) were collected from Pollution Control Board websites. The meteorological data namely relative humidity, rainfall, wind speed, wind direction, cloud octa (% of sky occupied by clouds), air pressure at station level and mean sea level were collected by request from the meteorological department. The data was collected and procured from the Meteorological Department for the three years, 2009 to 2011. Statistical tools were used to remove outliers and provide missing values. The data of first two years was used for training the neural network, and the third year data was used for testing and validation.

### 4.3 Hypothesis Development and Testing

The following null hypotheses are formulated and tested on data collected.

- **H$_0$$_1$**: The relative humidity will not affect the pollution level SO$_2$
- **H$_0$$_2$**: The relative humidity will not affect the pollution level NO$_x$
- **H$_0$$_3$**: The relative humidity will not affect the pollution level RSPM
- **H$_0$$_4$**: The wind direction will not affect the pollution level SO$_2$
- **H$_0$$_5$**: The wind direction will not affect the pollution level NO$_x$
- **H$_0$$_6$**: The wind direction will not affect the pollution level RSPM
- **H$_0$$_7$**: The wind speed will not affect the pollution level SO$_2$
- **H$_0$$_8$**: The wind speed will not affect the pollution level NO$_x$
- **H$_0$$_9$**: The wind speed will not affect the pollution level RSPM
- **H$_0$$_{10}$**: The cloud pattern will not affect the pollution level SO$_2$
- **H$_0$$_{11}$**: The cloud pattern will not affect the pollution level NO$_x$
- **H$_0$$_{12}$**: The cloud pattern will not affect the pollution level RSPM
- **H$_0$$_{13}$**: The rainfall will not affect the pollution level SO$_2$
- **H$_0$$_{14}$**: The rainfall will not affect the pollution level NO$_x$
\( \text{H}_{0_{15}} \): The rainfall will not affect the pollution level RSPM
\( \text{H}_{0_{16}} \): The maximum temperature will not affect the pollution level \( \text{SO}_2 \)
\( \text{H}_{0_{17}} \): The maximum temperature will not affect the pollution level \( \text{NO}_x \)
\( \text{H}_{0_{18}} \): The maximum temperature will not affect the pollution level RSPM
\( \text{H}_{0_{19}} \): The minimum temperature will not affect the pollution level \( \text{SO}_2 \)
\( \text{H}_{0_{20}} \): The minimum temperature will not affect the pollution level \( \text{NO}_x \)
\( \text{H}_{0_{21}} \): The minimum temperature will not affect the pollution level RSPM
\( \text{H}_{0_{22}} \): The air pressure will not affect the pollution level \( \text{SO}_2 \)
\( \text{H}_{0_{23}} \): The air pressure will not affect the pollution level \( \text{NO}_x \)
\( \text{H}_{0_{24}} \): The air pressure will not affect the pollution level RSPM

### 4.4 ANN Model for Prediction of Air Pollution

ANN model was developed for data mining and prediction of air pollution. The proposed model was trained, tested and validated with the case study data for supervised learning. The ANN model can be continuously improved to reduce variation in forecast as follows. The forecasts of the pollutants are reentered into the ANN model in order to improve the learning process and increase the efficiency in forecasting. The entire process is a four phase development consisting of initial research, model development, implementation, and testing. The predictive model is applied for any pollutant at any location.

### 4.5 Kriging Model for Prediction of Air Pollution

Kriging model was developed for predicting air pollution in nearby areas. A model based on R-software was developed and implemented for this purpose. This model takes a number of input data, including a field of observed data, the estimated range, the resolution of the estimated range, etc. and generates as output an estimated value and error variance. The user interface takes latitude and longitude of new location and generates a kriging value. The software can be used to assess the variations of meteorological, air pollution and related data at each monitoring
stations.

The data from Maharashtra Pollution Control Board (MPCB) and Meteorological departments (MET) are fed into a database from which it is taken as input to the ANN model. The ANN model is then run to train using selected test data and validation of the parameters carried out. The variables are predicted and the values are sent back to GIS as a feedback. The predicted values of the air pollutants are then feed into the Kriging model which interpolates the values of the air pollutants are the unmonitored locations in Mumbai. Figure 4.1 shows the proposed integrated model for prediction of pollutants in the atmosphere.

![Proposed Conceptual Model for Forecasting Air Pollutants](image)

Figure 4.1: Proposed Conceptual Model for Forecasting Air Pollutants

### 4.6 Proposed Algorithm

1. Initialize the network using random numbers with weights to be small random numbers between 1 and +1
2. Apply the input pattern
3. Predict the output i.e. target and apply it to the validation Set
4. Calculate the error as
   \[ Error_B = Output_B(1 - Output_B)(Target_BOutput_B) \]
5. Change the weight. Let \( W + AB \) be the new (trained) weight and \( W_{AB} \) be the initial weight.
\[ W + AB = WAB + (\text{ErrorB} \ast \text{OutputA}) \]

6. Calculate the Errors for the hidden layer neurons

7. Repeating this method until the network is trained

8. Forecast the output parameters for input parameters and create a matrix of forecasted values for the next three days

9. For each day

9.1. Get the forecasted output parameters.

9.2. Find the valid input points

9.3. Determine the distances between all valid input points (n) and find the semi-variogram value for these distances

9.4. For each of the output pixel

9.5. Determine the distances towards all input points, and find the semi-variogram value for these distances

9.6. Calculate the weight factors (vector w):

9.7. Calculate the estimated or predicted values for this output pixel

9.8. Calculate the error variance and standard error for this output pixel

**Formulae to calculate weight factors**

The Kriging weight factors of n valid input points \( i(i = 1, .., n) \) are found by solving the Eq.4.1

\[
[C] \ast [w] = [D] \tag{4.1}
\]

Eq. 4.1 can be expanded and rewritten in the matrix form as in Eq.4.2

\[
\begin{pmatrix}
0 & \gamma(h_{12}) & \gamma(h_{13}) & \ldots & \gamma(h_{1n}) & 1 \\
\gamma(h_{21}) & 0 & \gamma(h_{23}) & \ldots & \gamma(h_{2n}) & 1 \\
\gamma(h_{31}) & \gamma(h_{32}) & 0 & \ldots & \gamma(h_{3n}) & 1 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
\gamma(h_{n1}) & \gamma(h_{n2}) & \gamma(h_{n3}) & \ldots & 0 & 1
\end{pmatrix}
\times
\begin{pmatrix}
W_1 \\
W_2 \\
W_3 \\
\ldots \\
W_n
\end{pmatrix}
= 
\begin{pmatrix}
\gamma(h_{p1}) \\
\gamma(h_{p2}) \\
\gamma(h_{p3}) \\
\ldots \\
\gamma(h_{pn})
\end{pmatrix} \tag{4.2}
\]
This matrix equation is a set of $n + 1$ simultaneous equations as in Eq. 4.3

$$\sum_i (w_i \gamma(h_{ik})) + \lambda = \gamma(h_{pi}) \text{ for } k = 1, \ldots, n$$  \hspace{1cm} (4.3)

where the weights add up to 1 as in Eq. 4.4

$$\sum_i w_i = 1 \hspace{1cm} (4.4)$$

where

- $h_{ik}$ is the distance between input point $i$ and input point $k$;
- $h_{pi}$ is the distance between the output pixel $p$ and input point $i$;
- $\gamma(h_{ik})$ is the value of the semi-variogram model for the distance $h_{ik}$, i.e. the semi-variogram value for the distance between input points $i$ and input point $k$;
- $\gamma(h_{pi})$ is the value of the semi-variogram model for the distance $h_{pi}$, i.e. the semi-variogram value for the distance between the output pixel $p$ and input point $i$;
- $w_i$ is a weight factor for input point $i$;
- $\lambda$ is a Lagrange multiplier, used to minimize possible estimation error;

Matrix $C$ contains the semi-variogram values for all combinations of valid input points that will make a contribution to the output pixel value. Vector $w$ contains the weight factors for all valid input points that will make a contribution to the output pixel value. Vector $D$ contains the semi-variogram for an output pixel and all combinations of valid input points. Equation 4.3 guarantees unbiasedness of the estimates. The solutions $w_i$ minimize the Kriging error variance $\sigma^2$. 

68
Formula to calculate an estimate or predicted value for an output pixel

The estimate for the output pixel as in Eq.4.5

\[
\hat{Z} = \sum_i w_i \ast Z_i
\]  \hspace{1cm} (4.5)

where

\(\hat{Z}\) is the estimate or predicted value for one output pixel to be calculated;

\(w_i\) is the weight factor for input point \(i\);

\(Z_i\) is the value of input point \(i\);

**Formulae to calculate error variance and standard error**

The error variance can be calculated using the Eq. 4.6

\[
\sigma^2 = \sum_i (w_i \ast \gamma(h_{pi})) + \lambda
\]  \hspace{1cm} (4.6)

The standard error or standard deviation is the square root of the error variance, thus:

\[
\sigma = \sqrt{\sum_i (w_i \ast \gamma(h_{pi})) + \lambda}
\]  \hspace{1cm} (4.7)

where

\(\sigma^2\) is the error variance for the output pixel estimate ;

\(\sigma\) is the standard error or the standard deviation of the output pixel estimate

\((h_{pi})\) is the distance between the output pixel \(p\) and input point \(i\)

\(\gamma(h_{pi})\) is the value of the semi-variogram model for the distance \(h_{pi}\), i.e. the semi-variogram value for the distance between the output pixel \(p\) and input point \(i\);

\(w_i\) is a weight factor for input point \(i\);
4.7 Architectural Framework

The Figure 4.2 below gives the Proposed architectural framework of the Model developed. It consists of a three layer architecture consisting of the database server, Application Server and the Client. The database server consists of the database which has integrated the data from the MPCB and MET Department. The application server is the combination of the ANN and Kriging implementation in order to forecast the air pollutants at monitored and unmonitored locations. Finally the last layer is the Web interface which disseminates the predicted values on the web browser.

![Figure 4.2: Proposed Architectural Framework for the Integrated Framework](image)

The proposed flow chart of the Air Pollution Prediction Model is as shown in the Figure 4.3. The data is obtained from the MET and the MPCB are integrated into the database. The data then passes through an Artificial Neural Network module to forecast the air pollutants at the monitored locations. The results of the Artificial Neural Network are then passed through a Kriging Module to forecast the air pollutants at the unmonitored locations.
Air pollution model is being prepared using past available data from the MET database. The model is trained, tested and validated using three years of MET data (2009 - 2011). The proposal is to analyze, mine the data and develop computer based modules as below

**Assessment Module** - The main objective of this module is to assess the variations of meteorological, air pollution and related data at each monitoring stations. This is used to develop various sub-modules which will be used to support the development module, e.g., meteorological sub-module, pollution content sub-module, etc.
Development Module - This module is used to develop the predictive models for any pollutant at any location. For each pollutant, a separate model may be developed based on statistics of the pollutant records of previous years.

Control Module - This module is used to estimate and control the pollution level of criteria pollutant. This module helps the planners/environmentalists to identify the required decisions which must be taken to achieve the goal.

User Interface - It consists of menu-based interface to help various planners and decision makers in efficient usage of the developed Decision Support System. All the modules should be well linked together within a user interface and should provide graphics, dialog boxes, spatial analysis and other required functions. The research is divided into two sections. In the first section, the levels and trends of air pollution in Mumbai and Navi Mumbai city has been studied. In the next section, an attempt has been made to establish a linkage between the level of air pollution and the various meteorological parameters. In this thesis, the focus is on the air pollutants and their safe levels and the trends in Mumbai and Navi Mumbai as well as the health conditions of the exposed people in the survey areas.

4.8 Tools and Techniques for Data Analysis

1. The statistical analysis of data, viz. tabulation and graphical analysis of data, regression analysis, and nonlinear modeling was done using MINITAB software.

2. The computational model for ANN was developed using Neural Network toolbox of MATLAB.

3. The kriging model for interpolation of the pollutant levels was done using R software.

4. A web interface for the user was developed using PHP. This was integrated with the ANN model and kriging model using C sharp language.
4.9 Scope of the Study

The universe of study is limited to Mumbai and Navi Mumbai. Pollution data was obtained from different monitoring locations in Mumbai and Navi Mumbai. The monitoring stations in Mumbai are Sion, Bandra and Mulund. The monitoring stations in Navi Mumbai are Vashi, Nerul, Airoli, Mahape and Rabale areas. The meteorological data was collected from the Meteorological Department in Santa Cruz. The functional scope is confined to offering meaningful suggestions for improving prediction with ANN and Kriging tools only.