Application of Advanced Mining Algorithms to Diagnose Spatial Data

A Synopsis Submitted to the S.N.D.T. Women’s University, Mumbai For the Award of the Degree of

DOCTOR OF PHILOSOPHY IN TECHNOLOGY

UNDER THE FACULTY OF COMPUTER SCIENCE

BY

SUHASINI VIJAYKUMAR KOTTUR

UNDER THE GUIDANCE OF

DR. S. S. MANTHA

PROFESSOR, MUMBAI UNIVERSITY

AND CHAIRMAN, AICTE, N. DELHI

SHREEMATI NATHIBAI DAMODAR THACKERSEY (S.N.D.T.)

WOMEN’S UNIVERSITY, MUMBAI, MAHARASHTRA, INDIA

NOVEMBER, 2014
DECLARATION

Certified that this dissertation / thesis titled “Application of Advanced Mining Algorithms to Diagnose Spatial Data” is a bonafide record of work done by Mrs. Suhasini Vijaykumar Kottur during the period 2008 to 2014 at the P.G. Department of Computer Science, S.N.D.T. University, Juhu campus, Mumbai.

The synopsis of the aforesaid thesis is hereby submitted in partial fulfillment of the requirements for the Ph.D. Degree in Technology under the Faculty of Computer Science.

Signature of Guide

Name of Guide Dr. S. S. Mantha

Date:

I declare that the form and content of the above mentioned thesis and synopsis are original and have not been submitted, in part or full, for any other degree/ diploma of this or any other University or Institution.

Signature of scholar

Name of Scholar

Date:
THESIS APPROVAL LETTER

Acad.RRC-29/2008-09/77 dt. 15 May, 2008
# Table of Contents

**CHAPTER 1** ............................................................................................................................................. 6

**INTRODUCTION** ....................................................................................................................................... 6

1.1 Introduction ............................................................................................................................................ 6
1.2 Data Mining of Spatial Data .................................................................................................................. 6
1.3 Spatial Data of Air Pollution ................................................................................................................ 6
1.4 ANN Based Prediction Models for Air Pollution Data ......................................................................... 7
1.5 Kriging Models for Spatial Air Pollution Data .................................................................................... 7
1.6 Organization of the Synopsis Report ................................................................................................. 8

**CHAPTER 2** ............................................................................................................................................... 9

**REVIEW OF LITERATURE** .................................................................................................................... 9

2.1 Introduction ........................................................................................................................................... 9
2.2 ANNs for Air Pollution Modelling ..................................................................................................... 9
2.3 Mining of Spatial Data .......................................................................................................................... 10
2.4 Spatial Data and Kriging ..................................................................................................................... 10

**CHAPTER 3** ............................................................................................................................................. 12

**RESEARCH METHODOLOGY** .............................................................................................................. 12

3.1 Introduction .......................................................................................................................................... 12
3.2 Motivation ............................................................................................................................................ 12
3.3 Problem Statement ............................................................................................................................ 12
3.4 Research Objectives .......................................................................................................................... 13
3.5 Research Questions ........................................................................................................................... 14
3.6 Research Methodology ....................................................................................................................... 14

3.6.1 Data Collection ............................................................................................................................... 15
3.6.2 Hypothesis Development and Testing ......................................................................................... 15
3.6.3 ANN Model for Prediction of Air Pollution ............................................................................... 16
3.6.4 Kriging Model for Prediction of Air Pollution ............................................................................ 17
3.6.5 Conceptual Model ....................................................................................................................... 17
3.6.6 Proposed Algorithm ..................................................................................................................... 17
3.6.7 Architectural Framework Model ................................................................................................. 20
3.6.8 Tools and Techniques for Data Analysis ..................................................................................... 21

3.7 Scope of the study ............................................................................................................................... 21
CHAPTER 1
INTRODUCTION

1.1 Introduction

In the last few years, data mining has been the subject of active research. This thesis presents the application of advanced data mining techniques to spatial data of air pollution obtained from various Pollution Control Boards (PCBs) and meteorological sources. The motivation for this research comes from the problems faced due to increased air pollution in recent years particularly in urban areas and the possible application of developments in computer science specifically in the domain of data mining tools and technologies to mitigate this practical problem. The research work presents the details of methodology used for forecasting and interpolation of pollution data. The results obtained using Artificial Neural Networks (ANNs) and kriging techniques on live case studies on air pollution are presented.

1.2 Data Mining of Spatial Data

Data mining is the process of discovering meaningful patterns by sifting through data using pattern recognition technologies such as neural networks and machine learning and genetic algorithms [28]. Data mining enables automatic discovery of patterns in the data. Spatial data refers to pollution or meteorological data after the inclusion of geographical reference. Mining of spatial data has attracted the attention of several researchers in the study of phenomena like forest fires, air pollution, droughts, flood, and other applications. The present research deals with application of data mining to spatial data of air pollution.

1.3 Spatial Data of Air Pollution

Air quality is deteriorating particularly in urban locations. The World Health Organization estimates that about two million people die prematurely every year as a result of air pollution, while many more suffer from breathing ailments, heart disease, lung infections and even cancer. In view of the health hazards posed by increasing air pollution, it will be useful to have a model that can predict the level of atmospheric pollutants [40]. The present research deals with two major areas in air pollution modelling: (i) forecasting of level of pollutants at known locations
using ANN based models, and (ii) interpolation of level of pollutants at nearby unknown locations using kriging techniques.

1.4 ANN Based Prediction Models for Air Pollution Data

Artificial intelligence based methodologies such as Artificial Neural Networks (ANNs) can help to predict the pollutants in complicated non-linear contexts. The predictive accuracy obtained by ANNs is often higher than that of other methods or human experts [1]. In the simplest ANN model, the effects of the synapses in biological nervous systems are captured by weightages assigned to various input signals and learning occurs by adjusting the weights according to the learning algorithm[27]. Current ANN models use a multi-stage configuration where the output of a stage can be used as feedback to adjust the weights at other stages. Several ANN architectures are now available to suit different situations. The present research work proposes and tests an ANN model using MATLAB for prediction of pollution data at known locations in Mumbai and Navi Mumbai.

1.5 Kriging Models for Spatial Air Pollution Data

In a spatial context, the estimation and prediction of the values of a variable at unsampled locations is often of interest. This is achieved by modelling of spatial pattern using spatial interpolation techniques of advanced data mining. Krige proposed an interpolation technique by using a system of linear equations based on previous knowledge of the degree of spatial dependence in the data [22]. Kriging is a set of linear regression equations that determine the best combination of weights to interpolate the data as in the inverse weight distance method by minimizing the variance as derived from the spatial covariance in the data. The present research work proposes a kriging model using the ‘R’ statistical package and the results of analysis are provided.
1.6 Organization of the Synopsis Report

The synopsis report is organized as follows. Chapter 1 introduces the context of the research work describing the air pollution problem and data mining tools for forecasting and kriging. It also presents a chapterwise organization of the thesis.

Chapter 2 provides a review of literature relevant to this research. The chapter provides a perspective of artificial neural networks, air pollution modelling and kriging. Chapter 3 describes the research methodology along with motivation and purpose of the study, problem statement, research objectives, assumptions and scope of the study.

Chapter 4 describes the case study and data collected. Chapter 5 presents the details of artificial intelligence technique, namely artificial neural network analysis for forecasting the pollution data and regression analysis for kriging and hypothesis testing of spatial data.

Chapter 6 presents the results and discussion of the findings of the study. Chapter 7 concludes the report with recommendations and scope for further work. A list of publications by the candidate and references are provided at the end.
CHAPTER 2
REVIEW OF LITERATURE

2.1 Introduction

This chapter traces the development of literature related to this research. In particular, a brief review of past research in artificial neural networks for air pollution modelling, mining of spatial data and kriging is presented so as to provide a perspective for the current work.

Air pollutants are namely carbon monoxide (CO), oxides of nitrogen (NOx), particulates matters and other gaseous oxide. Air pollution modelling uses mathematical and statistical techniques for trend analysis or forecasting. Trend in non-spatial data refers to a pattern observed over time periods [5]. Trend in spatial data is usually a pattern of change in some non-spatial attribute while moving away from that location. Analysis of trend in spatial data often requires regression analysis and related statistical analysis methods. The regression analysis produces the confidence for the discovered trend between the observed and the predicted values.

2.2 ANNs for Air Pollution Modelling

ANNs can be treated as one of the multi-variate, non-linear and non-parametric statistical methods ([5], [6], [7]). One major application area of ANNs is forecasting [4] since ANNs are best suited for problems where the knowledge required for solutions is difficult to specify but for which there are enough data or observations available[10]. Hu [41] used the Widrow’s adaptive linear network to weather forecasting. Due to the lack of a training algorithm for general multilayer networks at the time, the research was quite limited. There was not much development in the use of ANNs for forecasting until 1986 [35] who introduced the back-propagation algorithm. Werbos first formulates the backpropagation and finds that ANNs trained with backpropagation outperform the traditional statistical methods such as regression and Box-Jenkins approaches.[14][15][38]

Lapedes and Farber ([23], [24]) designed the feedforward ANNs that can accurately mimic and predict such nonlinear systems. Their results show that ANNs can be used for modeling and forecasting nonlinear time series with high accuracy. Jones et al. [21] improved the neural network approach by one dimensional Newton’s method to train the network

9
Zhang et al. [41] provide a general summary of ANNs in forecasting, providing guidelines for neural network modelling, a general paradigm of ANNs used for forecasting, modelling issues of ANNs in forecasting and relative performance of ANNs over traditional statistical methods.

Feedforward ANNs trained with the backpropagation have become popular and useful for modeling air pollution[30]. Multi-layer ANNs have been used to forecast Ozone [10] [15][17], Sulphur dioxide [8], NO₂ [19] and particulate matter [32] in the air. Multi-layer Perceptron Artificial Neural Network (MPNN) and Kohonen Neural Network (KNN) are the main ANNs that cover a variety of air pollution and meteorological modeling applications [11].

### 2.3 Mining of Spatial Data

Mining of spatial data offers several challenges on account of the volume and complexity of relationships between data. Such relationships influence the data through proximity of regions and the commonality of causes of variations. Mining of spatial data requires a combination of different methods, such as computational, statistical and visual. Typical tasks in mining of spatial data include: (i) data processing (filter, aggregate, or transform data), (ii) prediction of data values, (iii) regression analysis (identify dependencies), (iv) supervised classification (into predefined categories), (v) clustering (unsupervised classification), (vi) analysis of linkages between data (based on attributes), (vii) data visualization (scatter plots, histogram plots, and complex plots). Statistical methods for classification include decision trees, maximum likelihood estimation (MLE), linear discriminant function (LDF) and nearest neighbour methods.

### 2.4 Spatial Data and Kriging

A major difference between classical and spatial statistics is that the independence of samples is generally not valid in spatial data. Spatial data tends to be highly self-correlated and one of the concerns in spatial data analysis is to detect such patterns. Another distinct property of spatial data called spatial heterogeneity implies that the variation in spatial data is a function of its location.

Kriging as an interpolation technique developed in geostatistics for optimal spatial prediction using statistical methods at unobserved locations from observed values of nearer locations [37]. Kriging tools constitute the fundamental tools in the field of spatial statistics, developed by the
founder of geostatistics, Georges Matheron for spatial prediction of ore resources. Kriging is a form of weighted moving average estimator. The weights are assigned on the basis of model fitted to variogram function which represents spatial structure in the variable. As the interpolation is based on the choice of sample weights, the weights are assigned based on the spatial autocorrelation statistics of the sampled data set and the determination of weights is related to the variogram model and the number of sampling locations considered.

McDonnell & Burrough [9] have shown that for applications in geosciences, when data is moderately sparse, kriging is the best interpolation technique available. Kriging techniques are used in oceanography [36] and meteorology, but it took longer to become widespread in ecology ([25], [34]). Holdaway [19] suggested that the climatic record at forest plot locations may be useful in studying how forests will respond to future climatic change. Bezzi and Vitti [6] evaluated the results of different kriging interpolation approaches. The literature presents applications of kriging to map background air pollution data as NO\textsubscript{2}, PM\textsubscript{10}, O\textsubscript{3}, SO\textsubscript{2}, CO [3].

Although there can be many performance measures for an ANN forecaster like modeling time and training time, the most important measure of ANN performance is the prediction accuracy it can achieve beyond the training data. There are a number of measures of accuracy in the forecasting literature and each has advantages and limitations [29]. The most frequently used are: the mean absolute deviation (MAD), the sum of squared error (SSE), the mean squared error (MSE), the root mean squared error (RMSE), and the mean absolute percentage error (MAPE). The performance of kriging is affected by the number and spatial distribution of monitoring stations.[18][20][33]
CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction

This chapter describes the overall purpose of the study including the motivation, problem statement, research objectives, research methodology, assumptions and scope of the study.

3.2 Motivation

India has made rapid strides in industrialization, and it is one of the ten most industrialized nations of the world. But this status has brought with it unwanted and unanticipated consequences such as unplanned urbanization, pollution and the risk of accidents.

Data mining provides an excellent tool for solving air pollution management problems. The necessity for predicting air pollution incidentally assumes more of a convenience when it is known that data sets are available for free of charge in MPCB data repositories and at a nominal cost from meteorological centres. Once an efficient tool is built for abstracting these large quantities of data sets and deriving useful knowledge from the same, this can help to detect the vulnerability of the exposed people in the survey areas. This will enable planners to utilize the information to improve the health conditions of these areas for effective air pollution mitigation and management.

3.3 Problem Statement

In order to be able to predict air pollution conditions, it is necessary to handle historical data sets of the parameters considered. Considering the vast amount of data available and to distinguish the pattern and extent of relationships for efficient and useful knowledge extraction, there is a need for using techniques such as data mining.

Much of the spatial data obtained are sparse in nature, for example, pollution levels at different locations. Thus, a method for obtaining a continuous data set from a sparse data set is a practically useful need. Kriging methods for interpolation recognize that spatial variation of any continuous attribute is often too irregular to be modeled by a simple, smooth mathematical function.
The present research thesis examines the air pollution data at different locations in Mumbai and Navi Mumbai and attempts to forecast and interpolate the values for different locations. More specifically, this thesis uses data mining tools and techniques of artificial intelligence such as artificial neural networks to forecast the air pollutants namely SO₂, NOx and RSPM in the monitored locations and kriging as a technique to estimate their values in the unmonitored areas of Mumbai and Navi Mumbai.

### 3.4 Research Objectives

The main objective of the research is to offer scientific analysis of pollution data. By adapting methods from machine learning to pollution data analysis, the state of the art in data mining of spatial data will be advanced. Thus, the main aim of the research is to develop a data mining technique for air pollutants pattern recognition through the use of available data sets. The following are some of the other objectives:

- To offer scientific analysis of pollution data.
- To advance the state of the art in data mining of spatial data by adapting methods from machine learning to pollution data analysis.
- To know the air pollutants in Mumbai and Navi Mumbai.
- To develop a model for estimating the air pollution level.
- To adapt and integrate methods for spatial pattern analysis using techniques from exploratory spatial data analysis, statistics and artificial intelligence.
- To demonstrate the application of data mining techniques using available data sets from different locations and different data sources as case study.
- To develop an application program to implement a kriging module.
- To identify the factors affecting the prediction of air pollution levels.
- To identify the limitations in implementation of tools for predicting of air pollution levels.
3.5 Research Questions

Research questions to be answered in order to achieve the objective are:

Q1) Which function can efficiently describe air pollution pattern in datasets?

Q2) What is the relation between air pollution pattern function parameters and air pollution characteristics like SO$_2$, NOx, RSPM and the general air quality index?

Q3) What is the relation between the meteorological parameters and the air pollutants?

Q4) What are the meteorological elements that affect the dispersion of the air pollutants in the geographical area?

Q5) What is the spatial and temporal pattern of air pollution in the study area?

3.6 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 analysed and modelled. 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, analysed and formatted as per the requirements of the data mining and kriging tools developed. Then these tools were used to generate the desired prediction 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.
3.6.1 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 & 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. SO2, NO2, SPM, Temperature, wind speed and direction, etc. at sites viz. traffic junctions, dumping sites, industrial, residential societies etc.

The data for air pollutants (SOx, NOx and RSPM) were collected from Pollution Control Board websites. The meteorological data (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. These data were collected 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.

3.6.2 Hypothesis Development and Testing

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

H01 The relative humidity will not affect the pollution level SOx
H02 The relative humidity will not affect the pollution level NOx
H03  The relative humidity will not affect the pollution level RSPM
H04  The wind direction will not affect the pollution level SOx
H05  The wind direction will not affect the pollution level NOx
H06  The wind direction will not affect the pollution level RSPM
H07  The wind speed will not affect the pollution level SOx
H08  The wind speed will not affect the pollution level NOx
H09  The wind speed will not affect the pollution level RSPM
H10  The cloud pattern will not affect the pollution level SOx
H11  The cloud pattern will not affect the pollution level NOx
H12  The cloud pattern will not affect the pollution level RSPM
H13  The rainfall will not affect the pollution level SOx
H14  The rainfall will not affect the pollution level NOx
H15  The rainfall will not affect the pollution level RSPM
H16  The maximum temperature will not affect the pollution level SOx
H17  The maximum temperature will not affect the pollution level NOx
H18  The maximum temperature will not affect the pollution level RSPM
H19  The minimum temperature will not affect the pollution level SOx
H20  The minimum temperature will not affect the pollution level NOx
H21  The minimum temperature will not affect the pollution level RSPM
H22  The air pressure will not affect the pollution level SOx
H23  The air pressure will not affect the pollution level NOx
H24  The air pressure will not affect the pollution level RSPM

3.6.3  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 to take place. 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.
3.6.4 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.

3.6.5 Conceptual Model

![Conceptual Model Diagram]

Figure 1. Framework for forecasting air pollutants

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 1. Shows the proposed integrated model for prediction of pollutants in the atmosphere.

3.6.6 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 Apply it to the validation Set.
4. Calculate the error \( \text{ErrorB} = \text{OutputB} \times (1 - \text{OutputB}) \times (\text{TargetB} - \text{OutputB}) \)

5. Change the weight. Let \( W^{+AB} \) be the new (trained) weight and \( W_{AB} \) be the initial weight.
   
   5.1.1. \( W^{+AB} = W_{AB} + (\text{ErrorB} \times \text{OutputA}) \)

6. Calculate the Errors for the hidden layer neurons \( \text{ErrorA} = \text{Output A} \times (1 - \text{Output A}) \times (\text{ErrorB} \times W_{AB} + \text{ErrorC} \times W_{AC}) \)

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

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

   9.4.2. Calculate the weight factors (vector \( w \)):

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

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

Formulæ to calculate weight factors:

The Kriging weight factors of \( n \) valid input points \( i \) \((i = 1, ..., n)\) are found by solving the following matrix equation:

\[
\begin{bmatrix}
\text{C}
\end{bmatrix}
\ast
\begin{bmatrix}
w
\end{bmatrix}
= \begin{bmatrix}
\text{D}
\end{bmatrix}
\tag{1}
\]

This matrix equation is a set of \( n + 1 \) simultaneous equations:
\[ S_i \left( w_i \ast g(h_{ik}) \right) + 1 = g(h_{pi}) \quad \text{for } k = 1, \ldots, n \]  
\[ S_i w_i = 1 \]  

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 \( i \)
- \( g(h_{ik}) \) 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 \( k \)
- \( g(h_{pi}) \) 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 \( i \)
- \( l \) is a Lagrange multiplier, used to minimize possible estimation error

Matrix \( C \) is the semi-variogram which has the values for all possible combinations of valid input points that contribute to the estimation of the output pixel value. Vector \( w \) is the matrix of weight factors for each of the valid input that contribute to the estimation of the output pixel value. Vector \( D \) thus contains the semi-variogram for an output pixel and all combinations of valid input points.

Equation (3) guarantees unbiasedness of the estimates. The solutions \( w_i \) minimize the Kriging error variance \( s^2 \).

**Formulae to calculate an estimate or predicted value for an output pixel:**

\[ \hat{Z} = S_i (w_i \ast Z_i) \]  

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 is calculated as:

\[ s^2 = S_i \left( w_i \ast g(h_{pi}) \right) + 1 \]  

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

\[ s = \sqrt{S_i \left( w_i \ast g(h_{pi}) \right) + 1} \]  

where:
\( s^2 \) is the error variance for the output pixel estimate
\( s \) 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 \( i \)
\( g(h_{pi}) \) 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 \( i \)
\( l \) is a Lagrange multiplier, used to minimize possible estimation error

### 3.6.7 Architectural Framework Model

The following Figure 3.1 gives the 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 the integrated data form 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 3.1: Architectural Framework](image-url)
3.6.8 Tools and Techniques for Data Analysis

1. The statistical analysis of data, viz. tabulation and graphical analysis of data, regression analysis, and non linear modelling 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# language.

3.7 Scope of the study

The universe of study is limited to Mumbai 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.
CHAPTER 4

CASE STUDY

4.1 Introduction

This chapter presents the details of the case study from Mumbai and Navi Mumbai used for pollution data. The data used in this research are daily ambient minimum and maximum air temperature, relative humidity, wind speed and wind direction, atmospheric pressures at station level and mean sea level, and daily concentration of NOx, SOx and RSPM at three locations in Mumbai (Sion, Bandra and Mulund) and five locations in Navi Mumbai (Vashi, Nerul, Airoli, Mahape and Rabale) for a period of 3 years from 2009 – 2011. This data was provided by regional centre of Meteorology Department (Santa Cruz) and the Maharashtra Pollution Control Board. Further, the month and date information was also provided as input. In some subtle way, the information about seasons gets conveyed to the model through the month and date information.

4.2 Pollutants Monitoring

Air pollution monitoring is important to know the baseline status of various pollutants. The air pollutants that are being regularly monitored by PCBs are Sulphur Dioxide (SO2), Oxides of Nitrogen (NOX) and Particulate Matter (SPM and RSPM). Additional air pollutants like Carbon Monoxide, Ammonia, Hydrogen Sulphide, Polycyclic Aromatic Hydrocarbons, Lead and Ozone are also being monitored at selected locations. The monitoring of such pollutants is carried out for 24 hours (4-hourly sampling for gaseous pollutants and 8-hourly sampling for particulate matter) with a frequency of twice a week, to have one hundred and four (104) observations in a year. The monitoring of meteorological parameters such as wind speed and wind direction, cloudiness, relative humidity (RH) and temperature are also integrated with the monitoring of air quality.

4.3 Air Pollution in Mumbai and Navi Mumbai

Mumbai is located along the western coast of India from 18 deg 53’ north to 19 deg 16’ north latitude and from 72 deg. East to 72 deg. 59’ east longitude. Mumbai consists of a peninsula originally composed of seven islets. According to World Health Organisation (WHO), Mumbai
is one of the top ten most polluted cities in the world. Process emissions and those from fuel consumption constitute the main sources of air pollution. Major air pollution sources include a giant fertilizer/chemical complex; two oil refineries and a thermal power plant, all based in Chembur, a suburb on the eastern coast of Mumbai [30].

4.4 Pollution Monitoring Stations

The geographical location and monitoring frequency of the 8 monitoring stations of Maharashtra Pollution Control Board in Mumbai and Navi Mumbai is given in Table 4.1. The monitoring of meteorological data is done at Santa Cruz.

Table 4.1 Location and Monitoring Frequency

<table>
<thead>
<tr>
<th>Location</th>
<th>Lat / Long</th>
<th>Type</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sion</td>
<td>19.0400° N, 72.8600° E</td>
<td>Residential</td>
<td>Continuous Monitoring</td>
</tr>
<tr>
<td>Bandra</td>
<td>19.0544° N, 72.8406° E</td>
<td>Residential</td>
<td>Continuous Monitoring</td>
</tr>
<tr>
<td>Mulund</td>
<td>19.1717° N, 72.9560° E</td>
<td>Residential</td>
<td>Continuous Monitoring</td>
</tr>
<tr>
<td>Vashi</td>
<td>19.0800° N, 73.0100° E</td>
<td>Residential</td>
<td>Two days in a week</td>
</tr>
<tr>
<td>Airoli</td>
<td>19.1550° N, 73.0070° E</td>
<td>Residential</td>
<td>Two days in a week</td>
</tr>
<tr>
<td>Nerul</td>
<td>19.0330° N, 73.0200° E</td>
<td>Residential</td>
<td>Two days in a week</td>
</tr>
<tr>
<td>Rabale</td>
<td>19.1394° N, 72.9962° E</td>
<td>Industrial</td>
<td>Two days in a week</td>
</tr>
<tr>
<td>Mahape</td>
<td>19.1178° N, 73.0269° E</td>
<td>Industrial</td>
<td>Two days in a week</td>
</tr>
</tbody>
</table>

The pollutant data and meteorological data from the above stations was obtained for the three-year period 2009-2011. The first two years data was used for learning and training and the third year data was used for validation.
CHAPTER 5
MODELLING AND ANALYSIS

5.1 Introduction

This chapter presents the details of regression analysis for hypothesis testing of spatial pollution data, artificial neural network analysis for forecasting the pollution data and kriging for predicting pollutants in nearby areas of Mumbai and Navi Mumbai.

5.2 Hypothesis Testing

The data that was obtained from the available sources namely the air pollution control board and the meteorological department was inconsistent and required some preprocessing and cleansing.

Table 5.1: Analysis and Results of Hypothesis Testing

<table>
<thead>
<tr>
<th>Meteorological Element</th>
<th>Air Pollutant</th>
<th>Karl Pearson Correlation Coefficient</th>
<th>P-Value</th>
<th>Hypothesis Accept/Reject</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative Humidity</td>
<td>Sox</td>
<td>-0.202</td>
<td>0.000</td>
<td>P &lt; 0.05 Reject H₀₁ Accept H₁₁</td>
</tr>
<tr>
<td></td>
<td>Nox</td>
<td>-0.241</td>
<td>0.000</td>
<td>P &lt; 0.05 Reject H₀₂ Accept H₁₂</td>
</tr>
<tr>
<td></td>
<td>RSPM</td>
<td>-0.473</td>
<td>0.000</td>
<td>P &lt; 0.05 Reject H₀₃ Accept H₁₃</td>
</tr>
<tr>
<td>Wind Direction</td>
<td>Sox</td>
<td>-0.094</td>
<td>0.000</td>
<td>P &lt; 0.05 Reject H₀₄ Accept H₁₄</td>
</tr>
<tr>
<td></td>
<td>Nox</td>
<td>-0.179</td>
<td>0.000</td>
<td>P &lt; 0.05 Reject H₀₅ Accept H₁₅</td>
</tr>
<tr>
<td></td>
<td>RSPM</td>
<td>-0.179</td>
<td>0.000</td>
<td>P &lt; 0.05 Reject H₀₆ Accept H₁₆</td>
</tr>
<tr>
<td>Wind Speed</td>
<td>Sox</td>
<td>-0.246</td>
<td>0.000</td>
<td>P &lt; 0.05 Reject H₀₇ Accept H₁₇</td>
</tr>
<tr>
<td></td>
<td>Nox</td>
<td>-0.135</td>
<td>0.000</td>
<td>P &lt; 0.05 Reject H₀₈ Accept H₁₈</td>
</tr>
<tr>
<td></td>
<td>RSPM</td>
<td>-0.321</td>
<td>0.000</td>
<td>P &lt; 0.05 Reject H₀₉ Accept H₁₉</td>
</tr>
<tr>
<td>Cloud Pattern</td>
<td>Sox</td>
<td>-0.200</td>
<td>0.000</td>
<td>P &lt; 0.05 Reject H₀₁₀ Accept H₁₁₀</td>
</tr>
<tr>
<td>Octa</td>
<td>Nox</td>
<td>-0.292</td>
<td>0.000</td>
<td>P &lt; 0.05 Reject H₀₁₁ Accept H₁₁₁</td>
</tr>
<tr>
<td></td>
<td>RSPM</td>
<td>-0.565</td>
<td>0.000</td>
<td>P &lt; 0.05 Reject H₀₁₂ Accept H₁₁₂</td>
</tr>
<tr>
<td>Rainfall</td>
<td>Sox</td>
<td>-0.089</td>
<td>0.000</td>
<td>P &lt; 0.05 Reject H₀₁₃ Accept H₁₁₃</td>
</tr>
<tr>
<td></td>
<td>Nox</td>
<td>-0.109</td>
<td>0.000</td>
<td>P &lt; 0.05 Reject H₀₁₄ Accept H₁₁₄</td>
</tr>
<tr>
<td></td>
<td>RSPM</td>
<td>-0.269</td>
<td>0.000</td>
<td>P &lt; 0.05 Reject H₀₁₅ Accept H₁₁₅</td>
</tr>
<tr>
<td>Maximum Temperature</td>
<td>Sox</td>
<td>0.041</td>
<td>0.006</td>
<td>P &lt; 0.05 Reject H₀₁₆ Accept H₁₁₆</td>
</tr>
<tr>
<td></td>
<td>Nox</td>
<td>0.024</td>
<td>0.0108</td>
<td>P &lt; 0.05 Reject H₀₁₇ Accept H₁₁₇</td>
</tr>
<tr>
<td></td>
<td>RSPM</td>
<td>0.105</td>
<td>0.000</td>
<td>P &lt; 0.05 Reject H₀₁₈ Accept H₁₁₈</td>
</tr>
<tr>
<td>Minimum Temperature</td>
<td>Sox</td>
<td>-0.176</td>
<td>0.000</td>
<td>P &lt; 0.05 Reject H₀₁₉ Accept H₁₁₉</td>
</tr>
<tr>
<td></td>
<td>Nox</td>
<td>-0.267</td>
<td>0.000</td>
<td>P &lt; 0.05 Reject H₀₂₀ Accept H₁₂₀</td>
</tr>
<tr>
<td></td>
<td>RSPM</td>
<td>-0.427</td>
<td>0.000</td>
<td>P &lt; 0.05 Reject H₀₂₁ Accept H₁₂¹</td>
</tr>
<tr>
<td>Air Pressure</td>
<td>Sox</td>
<td>0.068</td>
<td>0.000</td>
<td>P &lt; 0.05 Reject H₀₂₂ Accept H₁₂₂</td>
</tr>
<tr>
<td></td>
<td>Nox</td>
<td>0.083</td>
<td>0.000</td>
<td>P &lt; 0.05 Reject H₀₂₃ Accept H₁₂₃</td>
</tr>
<tr>
<td></td>
<td>RSPM</td>
<td>0.135</td>
<td>0.000</td>
<td>P &lt; 0.05 Reject H₀₂₄ Accept H₁₂₄</td>
</tr>
</tbody>
</table>
Table 5.1 shows the analysis and results of testing the hypotheses shown in Chapter 3, Section 3.6.2. The table shows that in each case, the p-value < 0.05. Hence, in each case the null hypothesis (meteorological factor will not affect pollutant level) as given in Section 3.6.2 is rejected. Therefore, alternate hypothesis, that is, meteorological factor will have an effect on pollutant level is accepted. The data set obtained from the meteorological department and the MPCB consisted of missing data in case of air pollutants. Since the data set was large enough the missing data was removed and eliminated from the data set. The same process was followed for the seven monitoring sites of Mumbai and Navi Mumbai.

5.3 ANN Model

The occurrence of high level of air pollution concentration is associated with certain well-defined meteorological conditions. Hence the idea is to choose the group of similar situations corresponding to similar air pollution levels out of all possible states of atmosphere and meteorological conditions. Such initial selection of the input data would facilitate the network learning process by restricting separate cases from one similarity group. The idea of forecasting using ANN is hence reduced to two-step approach: (a) classification of meteorological and air quality (AQ) conditions, and (b) determination of air pollution pattern in a given meteorological situation using models for AQ forecasts. [16][24].

ANN modeling basically involves three stages, namely the data preparation and sampling, the training process and the validation and testing. Feed-forward back propagation neural network have been applied in this study. Data recorded in the year 2009, 2010 and 2011 by the meteorology department at Mumbai and the MPCB were used. Data from the seven monitoring stations of Mumbai and Navi Mumbai were used namely Bandra, Sion, Mulund, Vashi, Airoli, Nerul, Rabale and Mahape. Before using the ANN model to perform the forecasting task, the model must be trained. The training step is the process of learning in the form of weights stored in the nodes. Using these weights, an ANN can carry out complex non-linear mappings from its input nodes to its output nodes. The data was divided into two sets which become the learning set for ANN training and testing set to verify the efficiency and correctness of the developed model. The Matlab© Neural Network Toolbox was used for development of the air quality prediction model, due to its ease and flexibility. The model gives as output the predictions for the desired number of days.
The training algorithm is used to find the weights that minimize the overall error measure such as the sum of squared errors (SSE) or mean squared errors (MSE). Hence the network training is actually an unconstrained non-linear minimization problem. Secondly, the test set given to the network still in the learning phase, by which the error evaluation is verified in order to effectively update the best thresholds and the weights. Finally, validation set, is to evaluate whether the model has effectively approximated the general function representative of the phenomenon, instead of learning the patterns uniquely.

A proper selection of the input and output parameters is essential in order to make the ANN learn with a fast convergence. Typically, more input data is better so as to make the model more comprehensive and the analysis more definite. This research work includes nine input parameters namely the month, relative humidity, rainfall, Wind Speed and the wind Direction, cloud octa, temperatures (maximum and minimum), and the pressure in the areas of Mumbai and Navi Mumbai. The latitude and the longitude are also considered in order to include the spatial parameters of the monitoring stations. Corresponding to each element in an input vector there is an input node in the network input layer. All the inputs and targets have to be normalized in order to make them contribute with the same influence to the neural network. The idea in the modeling approach is to design a comprehensive model for the three day prediction of the three air pollutants for each of the seven monitoring stations. Figure 5.1 represents the input output model.

![Figure 5.1: Inputs and outputs of RBFNN model for prediction of Air Pollutants SOx, NOx and RSPM](image-url)
5.4 Kriging

Kriging is a method of interpolation which predicts unknown values from data observed at known locations. A method to express the spatial variation that minimizes the error of predicted values estimated by spatial distribution is required. The effectiveness of the kriging depends on how well the selected model fits the data. So the modelling of semivariograms and covariance functions should be such that it fits the semivariogram or covariance curve to the selected empirical data and the model can be used for interpolation of unmeasured variables. Semivariogram can be thought of as a dissimilarity function as the variance of the difference increases with distance whereas covariance acts as a similarity function where the variance of the difference decreases with distance.

The Kriging methodology consists of two stages: Firstly, semivariogram and covariance functions, and estimation of the statistical dependence values are obtained, because both measure the strength of statistical correlation as a function of distance. Secondly, prediction using generalized linear regression techniques (kriging) of unknown values is done. Thus, a kriged estimate is a weighted linear combination of the known sample values around the point to be estimated. Kriging allows the user to derive weights that result in optimal and unbiased estimates. It attempts to minimize the error variance and set the mean of the prediction errors to zero so that there are no over- or under-estimates. Figure 5.2 shows the Kriging Model.

![Figure 5.2 Kriging Model](image-url)
The R Project software is used to work with kriging using the geoR package. As a first step the locations at which observed values are available are plotted. A grid for which predicted values are desired is set up. A function takes sequences of values for horizontal and vertical coordinates and sets up the entire grid. Then calculate and plot the variogram (spherical or wave variograms) using the latitude and longitude of the stations. Then use ordinary kriging predictions for r values on our prediction grid.
6.1 Introduction

This chapter presents the details of regression analysis for hypothesis testing of spatial pollution data, artificial neural network analysis for forecasting the pollution data and kriging for predicting pollutants in nearby areas of Mumbai and Navi Mumbai.

6.2 Hypothesis Testing

Table 6.1 shows the sample size after removal of outliers.

<table>
<thead>
<tr>
<th>Stations</th>
<th>Bandra</th>
<th>Sion</th>
<th>Mulund</th>
<th>Vashi</th>
<th>Airoli</th>
<th>Nerul</th>
<th>Mahape</th>
<th>Rabale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total sample values</td>
<td>1030</td>
<td>603</td>
<td>579</td>
<td>1067</td>
<td>1079</td>
<td>299</td>
<td>259</td>
<td>299</td>
</tr>
<tr>
<td>Total sample values after removal of outliers</td>
<td>992</td>
<td>586</td>
<td>486</td>
<td>794</td>
<td>896</td>
<td>292</td>
<td>259</td>
<td>299</td>
</tr>
</tbody>
</table>

The data set is used for regression model building with the help of the Minitab Software. The regression model was first developed for the three monitoring stations namely Bandra, Vashi and Airoli. In each case, the p-value < 0.05. Hence, in each case the null hypothesis (meteorological factor will not affect pollutant level) as given in Section 3.6.2 is rejected. Therefore, alternate hypothesis, that is, meteorological factor will have an effect on pollutant level is accepted. The data set obtained from the meteorological department and the MPCB consisted of missing data in case of air pollutants. Since the data set was large enough the missing data was removed and eliminated from the data set. The same process was followed for the seven monitoring sites of Mumbai and Navi Mumbai.

A thorough analysis of the data using statistical techniques of descriptive statistics, and multicollinearity using variance inflation factors and autocorrelation. It is found that there is high correlation between the pressures measured at the station level and mean sea level. Hence only one pressure value is used for the further study. A thorough testing of hypothesis using the p values at 95 % confidence interval all the null hypothesis indicating the effect of the various meteorological parameters on the air pollutants are accepted concluding that the meteorological
parameters humidity, wind speed, cloud octa, rainfall, temperature and the pressure affect the concentrations of the air pollutants in the atmosphere.

A regression model was built for the air pollutants SOX, Nox and RSPM, but it was found that the linear regression model does not give a good fit and it does not improve further by introducing non linearity through squares and cubic effects as well. Hence the researcher has decided to use the artificial neural networks for building the predictor model for SOx, NOx and RSPM.

6.3 ANN Model

The ANN output shows that the predicted values are more closer to the actual values as compared to the results from the regression results. This is indicated by the R value as shown in the Figure 6.1.

The regression values obtained used linear regression model were less than 0.4 thus indicating that the relationship between the meteorological parameters and the air pollutants is non linear. Regression analysis for the ANN model was performed to check the correlation between the actual and predicted results. As shown in Figure. 6.1, a very good fit between the observed and predicted results was obtained as indicated by R value of 0.925. Figure. 6.2 shows the reduction in MSE values with increase in epoch. As an example, Table 6.3 gives the predictions for coming 9 days for three pollutants. The high R values show that the desired value of fit is obtained between predicted and observed.
Figure. 6.1: R values for regression analysis of training, validation, testing and all data

<table>
<thead>
<tr>
<th>Location</th>
<th>Pollutant</th>
<th>R-Value using Linear Regression</th>
<th>R-Value using ANN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bandra</td>
<td>Sox</td>
<td>0.08</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>NOx</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RSPM</td>
<td>0.63</td>
<td></td>
</tr>
<tr>
<td>Sion</td>
<td>Sox</td>
<td>0.30</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>NOx</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RSPM</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>Mulund</td>
<td>Sox</td>
<td>0.09</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>NOx</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RSPM</td>
<td>0.43</td>
<td></td>
</tr>
<tr>
<td>Vashi</td>
<td>Sox</td>
<td>0.11</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>NOx</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RSPM</td>
<td>0.64</td>
<td></td>
</tr>
<tr>
<td>Nerul</td>
<td>Sox</td>
<td>0.13</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>NOx</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RSPM</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>Airoli</td>
<td>Sox</td>
<td>0.12</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>NOx</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RSPM</td>
<td>0.53</td>
<td></td>
</tr>
<tr>
<td>Rabale</td>
<td>Sox</td>
<td>0.11</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>NOx</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RSPM</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>Mahape</td>
<td>Sox</td>
<td>0.16</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>NOx</td>
<td>0.24</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RSPM</td>
<td>0.35</td>
<td></td>
</tr>
</tbody>
</table>
Figure 6.2: MSE for training, validation, testing

The Predicted values for the three air Pollutants using the Artificial Neural Networks is an Table 6.3

Table 6.3: Forecasted values of the three pollutants at monitored locations for the next three days

<table>
<thead>
<tr>
<th>Sr No</th>
<th>Location</th>
<th>Sox</th>
<th>Nox</th>
<th>RSPM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Day 1</td>
<td>Day 2</td>
<td>Day 3</td>
</tr>
<tr>
<td>1</td>
<td>Bandra</td>
<td>19.69</td>
<td>17.94</td>
<td>17.02</td>
</tr>
<tr>
<td>2</td>
<td>Sion</td>
<td>25.19</td>
<td>21.05</td>
<td>25.58</td>
</tr>
<tr>
<td>3</td>
<td>Mulund</td>
<td>52.22</td>
<td>52.17</td>
<td>50.01</td>
</tr>
<tr>
<td>4</td>
<td>Vashi</td>
<td>27.22</td>
<td>27.84</td>
<td>22.63</td>
</tr>
<tr>
<td>5</td>
<td>Nerul</td>
<td>13.50</td>
<td>14.37</td>
<td>13.90</td>
</tr>
<tr>
<td>6</td>
<td>Airoli</td>
<td>24.59</td>
<td>33.09</td>
<td>21.84</td>
</tr>
<tr>
<td>7</td>
<td>Rabale</td>
<td>20.24</td>
<td>17.69</td>
<td>22.49</td>
</tr>
<tr>
<td>8</td>
<td>Mahape</td>
<td>23.28</td>
<td>24.35</td>
<td>16.63</td>
</tr>
</tbody>
</table>

6.4 Kriging

The Krigging model shows the values of the air pollutants at unmonitored locations. The unmonitored points were selected separately in Mumbai and navi Mumbai areas. The observed locations are used to set up a prediction grid. The ordinary kriging function using the model with constant mean and global neighbourhood was used to produce the kriging predictions at 10
locations in Mumbai and 5 locations which were selected separately in Mumbai and Navi Mumabi. The locations selected in Mumbai and Navi Mumbai are shown in Table 6.4

Table 6.4 : Unmonitored Locations Mumbai and Navi Mumbai

<table>
<thead>
<tr>
<th>Sr No</th>
<th>Area</th>
<th>Location</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mumbai</td>
<td>Borivali</td>
<td>19.2300</td>
<td>72.8600</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Malad</td>
<td>19.1861</td>
<td>72.8486</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Andheri</td>
<td>19.1190</td>
<td>72.8470</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Sakinaka</td>
<td>19.0881</td>
<td>72.8907</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>Dadar</td>
<td>19.0180</td>
<td>72.8448</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Sion</td>
<td>19.0400</td>
<td>72.8600</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>Girgaon</td>
<td>18.9530</td>
<td>72.8130</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>Colaba</td>
<td>18.9750</td>
<td>72.8258</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>Chembur</td>
<td>19.0587</td>
<td>72.8997</td>
</tr>
<tr>
<td>10</td>
<td>Navi Mumbai</td>
<td>Ghansoli</td>
<td>19.7000</td>
<td>72.5900</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>Mumbra</td>
<td>19.1767</td>
<td>73.0222</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>Kopar Khairane</td>
<td>19.1031</td>
<td>73.0106</td>
</tr>
</tbody>
</table>

The kriging model was deployed for obtaining the values at the above said locations for the next three days and the results obtained are tabulated in Table 6.4

Table 6.4 : Predicted values using Kriging at the unmonitored locations for the next three days

<table>
<thead>
<tr>
<th>Sr No</th>
<th>Area</th>
<th>Location</th>
<th>Sox Day 1</th>
<th>Sox Day 2</th>
<th>Sox Day 3</th>
<th>Nox Day 1</th>
<th>Nox Day 2</th>
<th>Nox Day 3</th>
<th>RSPM Day 1</th>
<th>RSPM Day 2</th>
<th>RSPM Day 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Borivali</td>
<td></td>
<td>41.80</td>
<td>41.57</td>
<td>39.05</td>
<td>30.822</td>
<td>27.21</td>
<td>24.61</td>
<td>183.40</td>
<td>158.14</td>
<td>120.23</td>
</tr>
<tr>
<td>2</td>
<td>Malad</td>
<td></td>
<td>37.97</td>
<td>37.44</td>
<td>35.27</td>
<td>36.36</td>
<td>31.10</td>
<td>29.01</td>
<td>186.81</td>
<td>158.86</td>
<td>124.16</td>
</tr>
<tr>
<td>3</td>
<td>Andheri</td>
<td></td>
<td>30.47</td>
<td>29.24</td>
<td>28.01</td>
<td>51.44</td>
<td>42.80</td>
<td>42.48</td>
<td>196.34</td>
<td>166.16</td>
<td>140.92</td>
</tr>
<tr>
<td>4</td>
<td>Sakinaka</td>
<td></td>
<td>32.58</td>
<td>30.45</td>
<td>31.27</td>
<td>87.05</td>
<td>77.87</td>
<td>84.24</td>
<td>220.55</td>
<td>219.38</td>
<td>221.67</td>
</tr>
<tr>
<td>5</td>
<td>Dadar</td>
<td></td>
<td>23.04</td>
<td>19.56</td>
<td>22.54</td>
<td>123.17</td>
<td>109.05</td>
<td>120.73</td>
<td>244.11</td>
<td>252.16</td>
<td>279.32</td>
</tr>
<tr>
<td>6</td>
<td>Sion</td>
<td></td>
<td>25.19</td>
<td>21.05</td>
<td>25.58</td>
<td>150.17</td>
<td>135.86</td>
<td>152.67</td>
<td>262.52</td>
<td>293.52</td>
<td>341.69</td>
</tr>
<tr>
<td>7</td>
<td>Girgaon</td>
<td></td>
<td>22.81</td>
<td>19.54</td>
<td>22.07</td>
<td>115.23</td>
<td>101.32</td>
<td>111.54</td>
<td>238.73</td>
<td>240.73</td>
<td>261.81</td>
</tr>
<tr>
<td>8</td>
<td>Colaba</td>
<td></td>
<td>22.88</td>
<td>19.54</td>
<td>22.21</td>
<td>117.62</td>
<td>103.64</td>
<td>114.30</td>
<td>240.34</td>
<td>244.15</td>
<td>267.06</td>
</tr>
<tr>
<td>9</td>
<td>Chembur</td>
<td></td>
<td>31.35</td>
<td>28.37</td>
<td>30.89</td>
<td>116.33</td>
<td>105.61</td>
<td>117.09</td>
<td>240.21</td>
<td>257.77</td>
<td>281.92</td>
</tr>
<tr>
<td>10</td>
<td>Ghansoli</td>
<td></td>
<td>23.73</td>
<td>30.67</td>
<td>21.88</td>
<td>101.60</td>
<td>108.28</td>
<td>107.80</td>
<td>132.19</td>
<td>195.88</td>
<td>165.04</td>
</tr>
<tr>
<td>11</td>
<td>Mumbra</td>
<td></td>
<td>24.11</td>
<td>31.78</td>
<td>20.88</td>
<td>100.13</td>
<td>108.12</td>
<td>109.12</td>
<td>142.80</td>
<td>193.79</td>
<td>140.94</td>
</tr>
<tr>
<td>12</td>
<td>Kopar Khairane</td>
<td></td>
<td>25.61</td>
<td>27.04</td>
<td>20.11</td>
<td>59.26</td>
<td>62.00</td>
<td>53.54</td>
<td>160.16</td>
<td>156.64</td>
<td>115.39</td>
</tr>
</tbody>
</table>
Figure 6.3 Kriging values of the Air Pollutant SOx for Unmintored locations of Mumbai

Figure 6.4 Kriging variance of the Air Pollutant SOx for Unmintored locations of Mumbai
The Figure 6.3 shows the results of kriging performed at the unmonitored locations for the air pollutant Sox in Mumbai and the Figure 6.4 shows the kriging variance and Figure 6.5 shows the perspective plot of predicted values for the air pollutant Sox in Mumbai.

**Figure 6.5**: Perspective plot of predicted values of the air pollutant SOx in Mumbai
CHAPTER 7
CONCLUSIONS

7.1 Major Findings

This chapter presents the major outcomes of the research carried out. The chapter provides brief results of regression analysis for hypothesis testing of spatial pollution data, artificial neural network analysis for forecasting the pollution data and kriging for predicting pollutants in nearby areas of Mumbai and Navi Mumbai.

1. All null hypothesis were rejected and hence the alternate hypothesis are accepted indicating that exists a relationship between meteorological parameters and the air pollutants. The linear regression model did not give a desired R value indicating that the relationship between the air pollutants and the meteorological parameters is non linear and cannot be modeled using the Regression Model.

2. ANN gave better R value and forecast as compared to simple regression models.

3. Kriging was able to predict pollutant values for unknown locations thus making it possible to estimate the values of the air pollutants at unmonitored locations

4. The model along with web interface thus is helpful and useful to the policy makers to take actions and mitigate the risks if any in case the air pollutant values reach a undesired value and the common man to know the Air Quality Index to take necessary actions

5. This prediction of air pollutants at monitored and unmonitored locations can be used to develop a Support System.

6. The Support System thus built will be of great use to the policy makers and Urban Development Authorities to take necessary actions.
7.2 Scope for Future Research

1. Genetic Algorithms are a promising artificial intelligence technique that are giving good results in the area of prediction and forecasting. The technique can be used to check if the results are better than using the ANN technique.[13][2]

2. Hybrid models can be applied by combining two techniques taking advantages of various techniques to further improve the forecasting of the air pollutants.
LIST OF PUBLICATIONS


References

[1] Ajith Abraham “Artificial Neural Networks”, Oklahoma State University, Stillwater, OK, US


[31] Okobiah, O, Mohanty, S.P. ; Kriging bootstrapped neural network training for fast and accurate process variation analysis , Quality Electronic Design (ISQED), 2014 15th International Symposium
[38] Werbos, P.J., 1974. Beyond regression: new tools for prediction and analysis in the 
market model. Neural Networks 1, 339–356.
the state of the art, International Journal of Forecasting, 14, 35-62.