Chapter 2

Review of Literature Survey

Developments in artificial intelligence, classical statistics and machine learning have evolved the developments in data mining in the past two decades closely supported by mature technologies like massive data collection, powerful multiprocessor computers and advanced data mining algorithms. This chapter traces the development of literature related to this research along these domains. In particular, research in air pollution, classical data mining, artificial neural networks, and kriging have been dealt with in detail so as to provide a perspective for the current work.

2.1 Development of Database Technology

Classical database technology began in the sixties with the developments in data collection and mechanisms for efficient creation of databases. Database technology has allowed for the development of methods for effectively organizing the data and optimizing the storage and query processing. Relational database systems developed in the seventies and its commercial adoption through the eighties allowed convenient data accessing using query languages and other user interfaces. Additional tools in data analysis are under development for deeper and intelligent data analysis, like data clustering, forecasting and kriging. Such tools can greatly contribute to decision making using large volumes of data. The limitations of available data mining tools offer scope for research into data mining tools.
2.2 Mining of Spatial Data

Classical studies in mining of non-spatial data often use statistical tools such as regression analysis, variance analysis, confidence intervals, cluster analysis and discriminant analysis for the study of data and their relationships. The foundations of most data mining algorithms are based on one or more of classical statistical tools. Classical statistical tools play the role of building blocks where-upon modern advanced statistical analysis is based. However, data mining is distinguished from other data analysis methods, based only on statistics by its ability to discover hidden structures in the 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, like 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), nearest neighbor methods, linear discriminant function (LDF), and several others.

Spatial data mining can be suitably classified into two explicitly as descriptive data mining and predictive data mining. Descriptive data mining describes a set of data in a concise manner presenting the overall properties of the data. Predictive data mining on the other hand constructs one or a set of models from the data attempting to predict the behavior of new data sets. The descriptive methods are the correlation analysis, clustering and associative rules whereas the predictive techniques are the classification, regression and the Markov models.
2.3 Association in Spatial Data


2.4 Clustering in Spatial Data

Clustering is basically the organization of data in classes as in classification. However in contrast to classification, in clustering, class labels are not known and the clustering algorithm discovers the acceptable classes. Since the classification is not based on the specified class labels clustering is known as unsupervised classification. Most of the approaches of clustering are based on the principle of maximizing the intra class similarity and minimizing the inter class similarity between the data objects. (Berkhin, 2006) Thus, clustering (Han, et al., 2001) is finding the distribution characteristics of spatial data, in a group of data objects based on the similarity, by making the data of each cluster have a high degree of similarity, and the data of different clusters as dissimilar as possible. The advantage of using cluster analysis is that no prior knowledge of distribution of the attributes is required. For different applications, a great number of spatial clustering algorithms have been developed (Jain and Dubes, 1988; Jain et al, 1999; Han ,Kamber 2000; Yeung et al., 2001,). Traditional clustering techniques can be hierarchical or partitioning based (Belkhin, 2006). Recently, there are several other clustering techniques for example, density-based methods and grid-based methods (Symons, 1981). K-means is the simplest clustering algorithm and the most popular one
which clusters the objects into K-predefined clusters based on the mean of the clusters (MacQueen, 1967). Kaufman and Rousseau, 1990) and Raymond et al, (1994, 2002) allocate data objects into k pre-defined clusters, and iteratively re-arrange objects with a goal to improve the quality of results obtained through clustering.

2.5 Trend in Spatial Data

Trend refers to a pattern observed over time periods as proposed by Berndt and Clifford (1996). Trend in spatial data is usually a pattern of change in some non-spatial attribute while moving away from a start object. Analysis of trend in spatial data often requires regression analysis and related statistical analysis methods. The movement is modeled using the neighborhood paths and regression analysis on the respective attribute from the starting point in order to describe the uniformity of change. The regression analysis produces the confidence for the discovered trend between the observed and the predicted values. Global trend considers all the neighborhood paths of the same length and all possible lengths and performs the regression once for each set of all the paths which have the same length. The algorithm returns a spatial trend having the maximum length. Local trend performs a regression first on the paths which have a length greater than the minimum length and the further neighborhood paths of these paths are considered only if the trend is significant. The algorithm returns back a set of spatial trends positive and negative both of which are significant. Berndt and Clifford (1996).

2.6 Artificial Intelligence (AI) Techniques

The traditional approach for air quality prediction uses mathematical and statistical techniques. In these techniques, initially a physical model was designed and then data is coded with mathematical differential equations. But such methods suffers from disadvantages like they provide limited accuracy as they were unable to predict the extreme points i.e. the pollution maximum and minimum cut-offs cannot be determined using such approach. Also, such methods were lengthy and
inefficient approach for better output prediction (Li and Hassan, 2009). But with the advancement in technology and research, an alternative to traditional methods has been proposed i.e. Artificial Intelligence (AI) Techniques can be used for prediction purposes. Among various types of soft computing techniques, the following are the major predictive model techniques.

- Artificial Neural Networks (ANN)
- Support Vector Machines (SVM)
- Fuzzy Logic (FL)
- Hidden Markov Model (HMM)
- Genetic Algorithm
- Particle Swarm Intelligence
- Hybrid soft computing techniques

2.6.1 Artificial Neural Network (ANN)

With the pioneering work of McCulloch and Pitts, Artificial Neural Networks (ANN) has its roots in wide interdisciplinary history from the early 1940s (Sharma et al., 2005). ANN emerged as a mechanism to mimic the humans brain processes (Rao, 2012). ANN is an intelligent system having the capability to not only learn and remember but also to create relations among the data. ANN is a connected network made up of simple processing units called the neurons, and a large number of weighted links. The acquired knowledge is stored in the neurons, and the signals and information are passed over the weighted links of the large network (Shruti et al., 2013). The prediction of air quality, effectively addressed by the prediction of various air pollutants like Sulphur, carbon monoxide, nitrogen, ozone, suspended particulate matter (SPM) by divided the data set into training, validation and verification further simulation using ANN (Rao, 2012). ANN was effectively addresses the prediction of Sulphur Dioxide distribution and the future concentration in the air by modeling the Sulphur Dioxide concentration and its distribution from the air pollution station (Shine, 2004).
2.6.2 Support Vector Machines (SVM)

Support vector machines (SVMs), also referred to as support vector networks are learning models which are supervised combined with associated learning algorithms. The SVMs try to recognize patterns by analyzing the data. These patterns are then used for classification and regression analysis. Based on the set of data taken as input, the support vector machine network predicts for each given input, the output from the two possible classes. The support Vector machines and models are thus a binary linear classifiers that are non-probabilistic. (Cortes et al., 1995). The SVM model provides a reliable alternative and advantages in times series data analysis for predicting the level of air pollutants (Li and Hassan, 2009). The possibility of applying SVM model prediction of ambient air pollutant is studied and has been projected as a promising approach in prediction of PM10 pollutant (Soawalak, 2011).

2.6.3 Fuzzy Logic

The term ”fuzzy logic” was introduced with the introduction of fuzzy set theories. The fuzzy set theory is a form of multi-valued logic and deals with reasoning that is not fixed and exact but approximate (Zadeh, 1965). Fuzzy Logic deals with reasoning and provides a better overview in the form of rules that defines all the conditions that are required for predicting the air pollution prediction (Li and Hassan, 2009). In sugarcane processing industry, fuzzy logic can be used to classify and quantify levels of pollution as poor, ordinary, very good and excellent. The Mamdani fuzzy inference system provides the results for prediction of the air quality in and around the sugarcane industry

2.6.4 Hidden-Markov Model (HMM)

A Hidden Markov model (HMM) is a traditional approach for prediction and time series analysis (Dong et al., 2009). A HMM is built on the associations between the attributes of specific data items and a data set. Hidden Markov Model (HMM), a probabilistic function of a Markov Chain, enables the prediction of PM2.5, using the meteorological measurements and its observation levels (Dong et al., 2009).
2.6.5 Genetic Algorithm

Genetic Algorithm is based on Darwins Theory (Shine et al., 2004). It begins with arbitrary created individual population. The fitness is evaluated and parents are selected from the individuals. Genetic Algorithms effectively addresses the change of the accumulation of the surrounding atmosphere and prediction of the thickness of the air pollutants (Shruti, 2013). Studies have shown that Genetic Algorithms are effectively applied in mining the best feature subset from a large database containing measurements of pollutant concentration. The mined results are then feed to a nearest neighbor algorithm to predict the daily maximum concentration for pollutants (Kalapanidas, 2002).

2.6.6 Particle Swarm Optimization (PSO)

Particle Swarm Intelligence is populated mining method that is similar to a school of flying birds. Particles are possible solutions to the problem that is to be solved. Each particle regulates its flying according to its own and its companions flying experience (Lu et al., 2002). A PSO can be implemented to train multi-layer perceptron thus effectively predicting the air quality parameters. The grade of atmospheric pollution and multi pollutants can be evaluated using the Particle swarm optimization algorithm which can be created using the characters of pellucid principle and physical explication (Wang, 2006).

2.6.7 Hybrid Soft Computing Techniques

The combination of more than one soft computing techniques forms Hybrid soft computing technique. A number of hybrid soft computing techniques are applied in assessment of air quality prediction efficiently. A hybrid soft computing technique with the combination of ANN along with Fuzzy logic or with HMM and even with other soft computing techniques can be very effective for air pollution prediction and time series analysis (Naveed et al., 2010). A neural network based on the Particle Swarm Optimization is an unique optimization algorithm to train the multi-layer perceptrons (Lu, 2002). The complex neural network with evolutionary algorithm, rtNEAT (real-time Neuro-Evolution of Augmenting Topologies); can
predict the air quality in a more effective manner (Naveed et al., 2010).

2.7 Artificial Neural Networks

Artificial Intelligence (AI) offers another powerful approach for data mining. In comparison to plain statistics, AI tools attempt to apply procedures that mimic human brain to data mining problems. AI tools are built upon heuristics and meta-heuristics that require immense computational processing power. Hence in the early years, only high end scientific or government data mining applications could afford the use of AI based data mining. However, when such computing became affordable in the eighties, more and more applications began to be developed for data mining using AI. Thus, the technology enablers for data mining using AI have evolved over several decades along with developments in computer technology. Continuous improvements in computer hardware in terms of processing power and memory, and ready availability of statistical and image processing software have lowered the cost and increased the accuracy of analysis.

During the nineties, AI tools began to incorporate ideas of machine learning into conventional AI and statistics to develop Artificial Neural Networks (ANNs). Machine learning can be thought of as an advancement of AI, because AI heuristics and advanced techniques of statistical analysis are combined together to improve the mining results. Machine learning can be easily imitated using ANN and readily implemented on modern computers. Machine learning allows computers to learn and make decisions on the data with little inputs or guidance from humans on the structure or patterns of data. While statistical tools are still required, advanced heuristics and algorithms allow the computer to interpret the structure of data in independent and interesting ways that was perhaps not possible to predict by humans (White, 1989).

2.7.1 Historical Background of ANN

McCulloch and Pitts pioneering work in 1943 proposed the first neuron model of computing elements. McCulloch (2009) and Minsky and Papert (1969) showed the limitations of a simple perceptron. The main drawback of the model could
not learn from the existing examples since the weights were fixed. Hebb put forward a learning method that adjusts the connection weight. The adjustments are based on the pre and post synaptic values of the variables. Hebb (1949) and Rosenblatt (1958) proposed models, where the weights were adjusted by the perceptron learning law.

Hopfield’s (1982) energy approach and the back-propagation learning algorithm for multilayer perceptrons by Werbos (Werbos, 1974, 1988) contributed to the major developments in ANN activity in 1982. This approach was then popularized by Rumelhart et al. (1986) and Anderson and Rosenfeld (1988).

## 2.7.2 Structure of ANN

An artificial neural network is prepared from interconnected elements which replicate the neurons learning process in the human brain. ANN is a closed loop structure as seen in Figure 2.1; the output of the model is matched with the target output. The error is then calculated, and reduced by using some learning rules, by adjusting the weights in the model in order to get the desired output.

![Figure 2.1: Typical ANN Structure](Source: Carlos, 2003)

One of the important advantages of the ANN is that the neural network behaves analogous to a black box model. Once the ANN is trained, it is quick at predicting the desired values. The implementation, training and the interpretation of the results is a challenge and is a limitation of the artificial neural networks.

An artificial neuron takes multiple inputs and gives out one output as seen in Figure 2.2. It operates in three modes namely learning mode, testing mode and validation mode. The neurons are modeled such that it consists of inputs
that are multiplied by weights and then calculated by a mathematical function. This calculation subsequently determines the activation of the neuron. Another function is then used to calculate the output of the artificial neuron. Higher the weight of the artificial neuron, stronger is the input with which is multiplied. The process of adjusting the weights is called learning or training the model.

By feeding an ANN with a data set of desired inputs and set of desired outputs the ANNs can be made to repeatedly adjust the weights of each interior link to model more accurately and determine the correct output for a given input. By exploiting the behavior of input and output of ANNs it can be trained with known data to predict the outcomes for new data. The ANNs thus trained can be used to predict the outcome of new independent input data. The data sets that are used in ANNs can now be divided into three distinct sets called training, testing and validation sets. The largest amongst the sets is the one for training and is used by neural network to study patterns present in the data. The testing set is then used to by the ANNs constructed to evaluate the generalization capability of the trained network. Finally, the validation set is used to check on the performance of the trained network.
2.7.3 The Network Architecture

The McCulloch-Pitts neural model using linear activation function can be demonstrated using the data flow diagram as shown in the Figure 2.3. This model is also called the linear threshold gate. The disadvantage of McCulloch-Pitts neuron model is that it only generates a binary output. Also weights and threshold values are also fixed. This restricts the application to diverse areas.

The perceptron model of Rosenblatt (1958) enhanced the McCulloch-Pitts model by incorporating Hebbian learning rule of adjusting weights Bose and Liang (1996). The perceptron model had an extra input to represent the bias. As shown in Figure 2.4, each input node is linked to an output node by an edge with a weight attached to it. The output of a perceptron is calculated by taking the weighted sum of the inputs by considering the training data set and subtracting a bias factor which is predetermined. It also uses an activation function on the result. The weights are then recalculated by comparing the output and the desired result of the training data. The process is continuously repeated until the performance of the model reaches a satisfactory level.

Multilayer ANNs can be seen as single layer networks of cascading groups that can achieve higher level of computational capabilities. The ANNs can be grouped into two categories namely feed-forward and recurrent networks based on...
the connection pattern. In the former the nodes on a particular layer can only interact with the nodes of the next layer. A recurrent neural network (Mandie and Chambers, 2001), allows for links to be made between one node to other nodes in the same or to even previous layers. Figure 2.5 shows a feed-forward multilayer neural network.

In this multilayer structure shown below, information is passed from the input nodes to the units in the first hidden layer. The outputs from the first hidden layer are then passed to the next layer, and so on. The overall response of the network to the activation pattern supplied by the source nodes in the input layer is constituted by the set of output signals of the neurons in the output layer of the network. The output units behavioral pattern depends on the activity of the hidden units and the weights between the hidden and output units.

The feed forward back propagation is a strong algorithm for allocating error responsibility through a multi-layer network in the form of the back propagation algorithm. The error at output units is calculated by the back propagation algorithm by finding a set of weights that solves a network learning problem. The error at the neurons in the layer directly preceding the output layer is a function of the errors on all units that use its output. The error effects in the output node are propagated backwards through the network, after every training iteration. This is done to combine a system similar to the non-linear multi-layer perceptron that is capable of making decisions with the objective error function (Rumelhart, 1986).
2.7.4 Activation Function

The activation function, also known as the transfer function, establishes the relationship between the inputs and the outputs of a node and a network. Some amount of non-linearity is introduced with the help of the activation function which required for most ANN applications. The simplicity of implementation has made the standard logistic or sigmoid function as the most common form of activation function used in the construction of ANNs. The sigmoid function as an activation function is shown in Figure 2.6.

Other continuous functions like such as hyperbolic tangent function, sine or cosine function, or a linear function can also be taken as the activation function (Chen and Chen, 1995) under certain conditions. The average behavior learning combined with logistic functions can also be used as suggested by Klimasauskas.
A hyperbolic tangent functions can be used as an activation function for problems which involve learning about deviations from the average which is suitable in applications related to forecasting problems.

### 2.7.5 Learning Algorithms

The three classes of learning algorithms are: supervised, unsupervised and hybrid. In supervised learning, the network is provided with a correct output for every input pattern. The errors or discrepancies between the desired and actual response for each node in the output layer are found via a forward pass. Weights are determined which make the network give responses close to the known correct responses. The best examples that are known of this technique occur in the back propagation algorithm, the delta rule, and the perceptron rule. In unsupervised learning also known as self-organization, the output unit is trained to react to patterns within the input. The network is adjusted through internal monitoring of performance. In this prototype, the system is statistically discovers the salient features of the input pattern. The combination so supervised and unsupervised learning is hybrid learning. Some of the weights are generally determined through supervised learning, while the remaining is obtained through unsupervised learning.
The Hebbian learning rule is based on the variation of synaptic connections amongst neurons. The idea behind the variation is that if two neurons are active concurrently, their interconnection should be strengthened. The perceptron learning rule is similar to the Hebb rule with the only difference being that the connection weights are not modified when the network responds correctly.

Back propagation networks overcome the limitations that single layer networks have by having more sophisticated learning rule (Sanchez, 1991). The direction of the error is estimated by taking the partial derivative of the error of the network with respect to each. The algorithm starts by applying the derivative to each of the weights, starting from backwards, first the output layer followed by the hidden layer and then the input layer. Hence, this algorithm is called the back propagation algorithm or the generalized delta rule (Haykin and Simon, 1999). The back propagation algorithm (Rumelhart et al., 1986) that is used in feed-forward ANNs uses supervised learning. The neurons send their signals in the forward direction and the errors are propagated backward direction (Cai et al., 2009).

Thus in these multi layered networks, the network error will decrease until it reaches the point of local minima. The idea here is to reduce the error generated which continues until the ANN learns the entire training data. The training starts initially with random weights and these are latter adjusted by means of the learning rule. It has been proven that the back propagation learning model with adequate hidden layers can used to approximate any nonlinear function to the required level of accuracy. This makes the back propagation learning neural network a good model for signal prediction and system modeling. The weight modification is controlled by the learning. Larger learning rate causes the local minimum to be too large, and the local minimum may be going beyond a threshold value constantly, resulting in oscillations. On the other hand, if the learning rate is too low, the number of iterations required may increase reducing the performance of the ANN. The choice of learning algorithm, learning rate and test data for a given problem depends on the complexity of the problem (Yetilmazel and Saral, 2007).

The problem with back propagation technique is that it stops once it has reached the local minima mainly due to the random initialization of weights. So, different minima are observed by training the algorithms with different initialization-
tion of weights. Several algorithms are available that can optimize the weights used and minimize the error function.

The LevenbergMarquardt algorithm is one such algorithm that gives a mathematical solution to the problem of minimizing nonlinear function given a set of parameters. The primary application of the LevenbergMarquardt algorithm is in the least squares curve fitting problem: given a set of m empirical datum pairs of independent and dependent variables, \((x_i, y_i)\), optimize the parameters \(\beta\) of the model curve \(f(x, \beta)\) so that the sum of the squares of the deviations becomes minimal.\(\text{(Eq.2.1)}\)

\[
S(\beta) = \sum_{n=1}^{m} [y_i - f(x_i, \beta)]^2
\]  

The LevenbergMarquardt algorithm is an iterative procedure. To start a minimization, the user has to provide an initial guess for the parameter vector, \(\beta\). In cases with only one minimum, an uninformed standard guess like \(\beta_T = (1, 1, ..., 1)\) will work ; in cases with multiple minima, the algorithm converges to the global minimum only if the initial guess is already somewhat close to the final solution.\(\text{(Eq.2.2)}\)

In each iteration step, the parameter vector, \(\beta\), is replaced by a new estimate, \(\beta + \delta\) To determine \(\delta\), the functions \(f(x_i, \beta + \delta)\) are approximated by their linearizations.

\[
f(x_i, \beta + \delta) \approx f(x_i, \beta) + J_i \delta
\]  

where

\[
J_i = \frac{\partial f(x_i, \beta)}{\partial \beta}
\]  

is the gradient (row-vector in this case) of \(f\) with respect to \(\beta\). The weight update vector \(\delta\) indicates by how much the network weights should be changed to reach a better solution. \(\lambda\), the damping factor would start with a small value say 0.1. The Levenberg-Marquardt equation is solved and the weights \(w\) are updated using the weight update factor \(\delta\). After the weights are updated the networks errors are recalculated for every entry in the training set. The iterations end when the new sum of squared errors means \(\lambda\) is decreased. If there is no reduction in
the sum of squared errors the new weights are rejected and the same method is repeated with a greater value for $\lambda$. This adjustment factor $v$ is used to adjust the value of $\lambda$, which is generally defined as 10. The value of $\lambda$ it is multiplied or divided by the value of $v$ depending on whether the value of $\lambda$ needs to increase or decrease. The process is done iteratively till the error reduces and the current iteration ends when the error is reduced. (Henry, 2011)

2.7.6 Testing and Validation

In the learning phase the error evaluation is verified to get the best weights and the thresholds. To generate a network that is robust and reliable, some noise and randomness are added to the training data so as to get the network trained and familiarized with noise and variability in real data set. ANNs can repeatedly infer the unseen part of a population correctly even though the sample data includes noisy information which is not possible to be generalized after learning. The validation data set is a new data set used to evaluate ANN and to check whether the model has efficiently approximated the general function.

2.7.7 Performance Measures

There are various performance measures for an ANN predictor namely in terms of the time required to model and train the ANN. Apart from these metrics the crucial and most important measure of ANN performance lies in the prediction reliability it can attain over and above the training data. An accuracy measure is usually defined in terms of the difference between the desired and the predicted value. There are different measures of accuracy in the forecasting, each having their own advantages and disadvantages (Makridakis, et al., 1983). 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).
2.7.8 Forecasting using ANN

Forecasting is one of the major application area of the ANNs (Sharda, 1994) since ANNs are best suited for problems where the knowledge required for solutions is difficult to particularize, but for which adequate observations and data is available. Thus the ANNs can be considered as a multivariate non-linear non-parametric statistical method (White, 1989; Ripley, 1993; and Cheng and Titterington, 1994). This approach towards developing a model having the ability to learn from the experience of training is helpful in case of many practical problems such as forecasting of air pollutants. As forecasting is made by predicting future behavior from examples of past behavior, this can be suitably justified as an ideal application for neural networks.

The earliest application of ANN for forecasting can be dated way back to 1964. Hu 1964, in his research, uses the adaptive linear network model developed by Widrows to forecast the weather. Since the training algorithm for general multilayer networks was not appropriately measured, the research was limited. There wasnt significant development in using ANNs for the purpose of forecasting until in the year 1986 when Werbos (1974, 1988) who introduced the back-propagation algorithm. The research first formulates the back propagation and finds that the ANNs that are trained using the back propagation method surpass the traditional statistical methods like regression and Box-Jenkins approaches (Rumelhart et al., 1986).

Lapedes and Farber (1987, 1988) reported one of the earliest applications of ANNs in forecasting and implemented it successfully. They designed the feed forward neural networks which accurately imitate and predict nonlinear systems similar to the ANN. The results obtained by them showed that ANNs can be accurately used for modeling and predicting nonlinear time series. Jones (1990) further improved the Lapedes and Farber models by one dimensional Newtons method to train the network. Poli and Jones (1994) further build a stochastic multilayer perceptron model with connections between units and noisy response functions using random values. It was Zhang et al. (1998) who provided the summary of the work in ANN forecasting by providing the guidelines for neural network modeling. They gave a general paradigm of the ANNs to be used for forecasting, modeling chal-
lenges of ANNs in forecasting and the relative performance of ANN as compared to the basic traditional statistical methods.

Feedforward ANN trained with the back propagation is a popular and convenient tool for modeling environmental systems, like air pollution. ANN technique has become quite popular for modeling air-quality data (Nejakoorki, 2011). The multilayer neural network technique has been used to forecast the ozone (Comrie, 1997; Gardner and Darling, 1996, 1999), the sulphur dioxide (Boznar, 1993), the NO$_2$ (Gardner and Darling, 1999) and the particulate matter (Parez and Triar, 2001) in the ambient environment. This is not only an intelligent approach but also a cost-effective one and hence has received much attention in environmental engineering. Multilayer perceptron artificial Neural Network (MPNN) and Kohonen neural network (KNN) are the main ANNs that can be used for the modeling of pollution and meteorological applications (Nejakoorki, 2011).

2.8 Spatial Data and Kriging

There is an increasing need for data mining applications that help in describing, predicting spatial patterns and explaining them (Ester et al., 1998, 1999; and Sanchez, 1991; Peuquet, 2000). A major differentiating point between classical and spatial statistics is that the samples cannot be considered to be mutual dependent in spatial data. Spatial data have a tendency to be highly self-correlated and one of the concerns in spatial data analysis is to detect such patterns. An additional distinguishing property of spatial data is the spatial heterogeneity which refers to the disparity in spatial data. This spatial heterogeneity in the spatial data which is a function of its location is measured using local methods of spatial autocorrelation.

2.8.1 History of Interpolation

Interpolation is the process of predicting the values of a variable of interest at unsampled locations based on measured values at some points within the area of interest (Burrough and McDonnel, 1998). Interpolation is a very successful tool in modeling changes in spatial data of any environmental system (Goodchild et al., 1993). Interpolation of spatial data enables the representation of a surface
and predicts values of other unknown areas. This prediction helps creating a continuous surface (Lam, 1983). Hence, any interpolation technique produces smoothed estimated values for unsampled locations.

Several interpolation methods are available, each having various advantages and limitations in the way the spatial structure is modeled. Based on the interpolator applied, interpolation techniques can be classified as global or local, exact or approximate, and deterministic or stochastic (Li and Heap, 2008).

Global interpolators normally make use of a single function on a complete dataset to interpolate the values for the entire study area. The resulting map of the interpolated data is usually a smooth surface. Changing one of the values affects the function and thus the predicted values. Local interpolators use single or multiple functions on subsets of the dataset within a fixed domain and change in one value will affect only neighboring locations (proximity polygons, inverse distance weighting, kriging). Thus, the interpolation function is applied locally only for a restricted number of locations. The resulting map of the interpolated data is still smooth but includes both global and local patterns.

Exact interpolators replicate the original values at the data points which are used to interpolate the unknown values (proximity polygons, inverse distance weighting and kriging). When approximate interpolators are applied at the sampling locations, the predicted values will not be the same as the observed ones (trend surface analysis).

Deterministic techniques to perform spatial interpolation provide no indication of the extent of possible errors while stochastic techniques provide probabilistic estimate. Some of the methods are the nearest-neighbor, the polynomial functions, radial basis functions (RBF) and inverse distance weighted (IDW) approaches (Isaaks and Srivastav, 1990; EPA, 2010, 2004; Johnson et al. 2001). The stochastic methods are the geostatistical approaches that include kriging (Beelen et al., 2009; Jourdon, 2009) and cokriging (Cressie, 1993). Deterministic interpolation techniques use the mathematical functions to calculate the values at unknown locations that are based on the amount of similarity (IDW) or the degree of smoothing (RBF) with respect to the neighboring data points. Geostatistical techniques use both the mathematical and statistical methods. These methods are based on the spatial autocorrelation that exists among the data points. Such methods predict
values at unknown locations and also provide probabilistic estimate the quality of
the interpolation. The spatial interpolation methodologies rely on the weighted
average approach are classified based on the difference in the approach that is
used to calculate the interpolation weights (W1). The Point based interpolation
methods estimate the discrete unsampled locations based on point observations of
the variable that is under study. The areal based schemes estimations are related
to the area or zone that is predefined. Global methods are based on the entire
data that is available of the variable under study and derive a global estimation
function for the area under study, whereas the local methods divide the area and
work independently within a space, using only a portion of the available data.
The methods that estimate the surface passing through the observed data point
are termed as interpolators, while the ones that use the available data and esti-
mate the surface and do not produce the observed values exactly the same are
the approximate interpolators. Stochastic interpolation methods incorporate the
idea of randomness taking into consideration the uncertainty associated with the
prediction, whereas the deterministic methods do not use the probability theory.
Depending on the degree of the estimation the methods of interpolation can be
categorized as gradual or abrupt.

2.8.2 Interpolation using Kriging

Kriging as an interpolation technique developed in geostatistics for optimal spatial
prediction at unobserved locations from observed values of nearer locations using
statistical methods (Waller, 2004). Kriging is a type of weighted moving average
estimator. The weights in kriging are assigned depending on a model fitted to
variogram function. The variogram function represents the spatial structure of
the variable under study. These interpolation schemes are named after D.G. Krige
who proposed the technique for interpolation (Krige, 1951). Krige (1966) proposed
a spatial interpolation method using a system of linear equations based on previous
knowledge of the degree of spatial dependence.

Interpolation consists of three involving defining the search area or area around
the point to be predicted, locating the observed data points within this area,
and assigning appropriate weights to each of the observed data points. As the
interpolation is based on the choice of sample weights, which are assigned based on the spatial autocorrelation of the sampled data set. The determination of weights is related to the variogram model and the number of sampling locations considered.

Kriging tools are the most important tools in the field of spatial statistics. Kriging was first developed to address specific needs in mining applications for spatial prediction of metal ore resources, such as interpolating from either point or block samples over a two-dimensional region, as well as predicting the values of the metal ore for a given three dimensional volume.

The Kriging technique makes the assumption that the distance dividing the sample points estimates a spatial correlation which can then be used to explain the variation in the surface. Kriging is a multi-step process that includes statistical analysis of data, description of the data, followed by the modeling using a variogram, creating the surface, and exploring a variance surface. Geostatistical methods like Kriging are built on statistical models that show the autocorrelation and statistical relationship among the measured points. This property of autocorrelation has the capacity of producing a prediction surface and providing a measure of the accuracy of the forecasts. Kriging can be done using Ordinary Kriging or Universal Kriging. Ordinary Kriging is the most widely used Kriging method and considers that the mean is constant and unknown. Universal Kriging assumes the trend prevalent in the data can be modeled using a polynomial that is a deterministic function. The deterministic function, polynomials that are subtracted from the original measures points and the autocorrelation is modeled from the random errors. Once the clear fit model to random errors is obtained, the polynomial is added back to the predictions to give meaningful results before making a prediction.

McDonnell and Burrough (1998) have presented that kriging is the best interpolation technique available in geosciences, when data is sparse. Kriging techniques are used in oceanography (Simard, 2003) and meteorology, and ecology (Legendre and Fortin, 1989; Rossi, 1992). Holdaway (1996) suggested the response of forest to future climatic changes can be studied by recording the climate at the forest plot locations. Bezzi and Vitti (2005) evaluated the results of different kriging interpolation approaches. The background air pollution data as NO$_2$, PM10, O$_3$,
SO$_2$, CO (Beelen et al., 2009) can also be interpolated using kriging. The performance of kriging is highly dependent on the number and spatial distribution of monitoring stations available which may be a problem in some cases.

2.8.3 Assumptions in Kriging

The characteristics of estimation are unbiasedness, stationarity and isotropy. Unbiased says that in the observation network every value has an equal probability of being included within the population. Kriging as a predictor does not demand normal distribution of the observed data. The observed values need to follow normal distribution to obtain probability maps and quantile for ordinary, simple, and universal kriging. If the data is normally distributed, kriging is the best predictor among all unbiased predictors, not only those that are weighted averages. The property of stationarity is required for drawing inferences from a model that describes the process of the spatial structure of data that are not samples at unknown locations. The assumption of stationarity is related to a property of the model or the process, and not to a property of the data. There are two types of stationarity in the spatial context namely mean and covariance. Mean stationarity is valid where the properties such as the mean and the variance of a process, are stationary (Haining, 1990; Burrough and MacDonnell, 1998). The second order stationarity is for covariance and intrinsic stationarity for semi-variograms. If the covariance is the same between any two points that are at the same distance and direction apart it is second order stationarity no matter which two points are selected. The covariance does not depend on the location of the two points but on the distance separating them (Fortin et al., 2003; Boots, 2002). Isotropy is the property of having the same characteristics in any direction whereas the opposite is anisotropy: when the system has changing characteristics in different directions.

Kriging is a comparatively fast interpolator and can be smoothed depending on the method. Kriging can get many outputs apart from the predictions maps, prediction errors and probabilities. A drawback is that it may require assessment to check for prediction errors or cross-validation (Tveito, 2007; Tveito et al., 2006).
2.8.4 Performance Measures in Kriging

The success of an interpolation method in predicting the values at unknown locations can be evaluated in using cross-validation, data splitting and calculation of the kriging variance. Cross-validation techniques check whether the model and its parameter values are reasonable. The cross validation works by removing one or more data locations and using the remaining data for predicting their associated values at the rest of the locations. Thus, the predicted value and the observed values can be compared to obtain useful information about the quality of the kriging model. In the data splitting technique, the dataset is split into two parts, one part is used in estimation and the other part is used for validation. This method is applicable when the numbers of observations are large and regularly sampled. The kriging variance is then estimated on the points where there are no observations providing a spatial view on the measure of success.

2.8.5 Kriging Models

Based on the system of linear equations, several kriging variants have been developed conserving the property of unbiased predictors

**Ordinary Kriging (OK)** is the simple form of kriging where the forecasting is a linear combination of the values measured. The spatial correlation amongst the data determines the weights and, is described by the variogram. Since the mean is unknown, the method assumes stationarity conditions. Although meteorological variables are often not stationary, this method is still used, often as part of residual kriging or indicator kriging.

**Simple Kriging (SK)** can be said to be Ordinary Kriging but where the mean is known. The mean is often demanding to derive, SK is slightly more dominant than OK.

**Block kriging (BK)** predicts the average values of the primary variable over a volume, surface, or segment, rather than at a point (Goovaerts, 1997). It extends the Ordinary Kriging by interpolating a block value by changing the point-to-point covariance with the point-to-block covariance (Wackernagel, 2003).

**Cokriging** is an interpolation technique that helps to estimate the primary variable if the distribution of a secondary variable is sampled more intensely than
the primary variable. Cokriging can also be said to be an extension of Ordinary Kriging using a covariance model or a multivariate variogram on multivariate data. Non-linear Kriging can be used when the data are non-normally distributed.

**Kriging with a Trend (KT) or Universal Kriging (UK)** is an extension of Ordinary Kriging that take account of the local trend in the neighborhood search space as a uniformly varying function of the coordinates. Universal Kriging forecasts the trend components within each search neighborhood space and then performs Simple Kriging on the corresponding residuals.

**Kriging with an External Drifts** is similar to Universal Kriging and includes the local trend within the neighborhood search space as a linear function of uniformly varying secondary variable (Goovaerts, 1997).

**Residual Kriging** is also called detrended kriging. In residual kriging the residuals generated from a regression that is previously fitted are interpolated using Ordinary Kriging and has wide applications in the area of meteorology (Tveito, 2007). Disjunctive Kriging is a non-linear process in where the dataset that follows a bivariate normal distribution is transformed using a series of additive functions.

**Indicator Kriging** is interpolation of a categorical variable that uses thresholds to create binary indicator values and then uses Ordinary Kriging for interpolation. Indicator kriging is a special version of disjunctive kriging and is not suitable for data having a trend (Tveito, 2007).

**Probability Kriging** is a non-linear method using indicator variables. The method is a form of cokriging using two variables, the first being the indicator and the second the original untransformed data (Tveito, 2007). It is powerful than indicator kriging but involves extra calculations since cross variances have to be fitted. This method is also not suitable for data having a trend.

**Stratified Kriging** is a global method and can be applied on complete datasets. But when there is sufficient information to categorize the area into meaningful sub-areas. Also there is sufficient data to compute the variogram for each sub-area; it is possible to carry out the interpolation in each separate area using stratified kriging (Burrough, 1998).

**Factorial Kriging (FK)** determines the origin of the value of a continuous characteristic (Goovaerts, 1997). It models the experimental semivariogram as a linear combination of a few basic structures to represent the different factors op-
erating at different scales (e.g., local and regional scales).

Dual Kriging interpolates the covariance values and not the data values to interpret the filtering characteristics of kriging (Goovaerts, 1997). It reduces the computational cost of kriging when combined with a global search neighborhood. It includes dual Simple Kriging, dual Ordinary Kriging, and dual Factorial Kriging.

Model based Kriging was proposed by Diggle et al. (1998). The method includes the linear kriging procedure but in a more general distributed structure that is similar to the structure of a generalized linear model. A Bayesian approach is implemented through the Markov chain Monte Carlo procedure, to predict arbitrary function of an unobserved hidden process while taking in consideration the uncertainty in the estimation of any model parameters (Ribeiro and Diggle, 2007). This technique demands a huge computation (Moyeed and Papritz, 2002), this method it is not applicable to large dataset.

2.9 Air Pollution

Clean air is one of the most essential components for the survival of living organisms including human beings. Air is a mixture of gases, minute solid and liquid particles. These substances originate from natural and anthropogenic sources such as industry, domestic activities, motor vehicles, shipping etc. and it can be either gaseous, liquids, or particles, in nature (Alias et al., 2007). The World Health Organization (WHO) reported in March 2014 that in 2012 around 7 million people expired because of exposure to air pollution, indoor as well as outdoor; which is one out of eight reasons for global deaths. It is confirmed that air pollution is now the worlds largest single environmental health risk. The large scale death due air pollution related problems was majorly found in low and middle revenue countries of Western Pacific Regions and South-East Asia. Thus, monitoring of air quality is an important task for environmental agencies, globally (WHO/UNEP, 1992; Lu and Wong, 2002).
2.9.1 Air Pollution and its Sources in India

Air pollution in India is mainly due fast industrialization, energy production, urbanization and commercialization. There has been a rise in population in urban areas from 62.4 million to almost a billion between 1951 and 2001, and its proportion has increased from 17.3% to 55.7%. About two-thirds of the urban population is situated in cities, 50% of which lives in 23 metropolitan areas with populations exceeding 1 million. This fast increase in urban population has led to the unintended urban development; increase in consumption patterns and larger demands for transport, energy, other infrastructure, leading to air pollution problems.

Vehicles, large number of industries and thermal power plants are the contributors to air pollution in urban areas, with vehicles being the major ones. India being a developing and agro-based country, many industrial areas and household activities involves burning of fossil fuels such as coal, wood and cow-dung that contain sulphur for domestic cooking purpose, generating lot of smoke. The dominant sources of air pollution in urban areas are often transport sector and in the rural areas, the domestic/residential wood combustion/cooking is also one of the major factors (Dalvi et al., 2006; Ghude et al., 2008, Sahu et al., 2008). The individual pollution episodes like wild fires, fire management burns and dust storms also contribute to air pollution. Particulate matter due to gaseous emissions from industries and auto exhausts are responsible for increasing uneasiness and number of respiratory diseases in urban centers.

Of particular concern is the vehicular emission as these are mainly at ground level and have a significant influence on the overall population. The other factors contributing to the growing vehicular pollution are the different types of engines and machinery used road conditions, traffic management systems, and rise in the density of traffic. There has been an increase in the number of motor vehicles from 0.3 million in 1951 to 37.2 million in 1997 (source: Ministry of Surface Transport, 2000). Based on this data, it can be approximated that the number of motor vehicles has increased above 50 million. Among the total increase in the vehicular traffic, 32% are concentrated in 23 metropolitan cities. At the all-India level, two-wheeled vehicles in the total number of motor vehicles increased from 9% in 1951 to 69% in 1997, and the share of buses declined from 11% to 1.3% during the
same period (source: Ministry of Surface Transport, 2000). This clearly shows that personal transport vehicles have increased tremendously. In 1997, personal vehicles two-wheeled vehicles and cars only constituted 78.5% of the total number of registered vehicles. The transport facilities in terms of cars and two wheelers are increasing which are helpful but simultaneously such facilities are giving rise to traffic issues and are also leading to an increase in the air pollution and need to be monitored regularly, these includes carbon monoxide (CO), oxides of nitrogen (NO\textsubscript{x}), particulates matters and other gaseous oxide. This kind of air pollution generates smog which subsequently leads to respiratory health problems and causes holes in the ozone layer, which increases the exposure of sun’s harmful rays.

The electricity generation capability in India has increased by about 55 times from 1.7 thousand MW to 93.3 thousand MW since 1950-51. The generation of electricity in India is done by a combination of nuclear, thermal, and hydro plants. The hydro-electric capacity also has changed significantly with the share of hydro-electricity in total capacity declining from 43% in 1970-71 to 24% in 1998-99, and the thermal power constituting about 74% of the total installed power generation capacity. This increase in the dependence on thermal energy has at the same time lead to numerous environmental problems.

The nature of particles from industrial sector depends on the materials used, products manufactured as well as the manufacturing process. Generally, in metallurgical plants, particle emission may contain heavy metal sublimates and silicates. If coal is used for cooking and reduction of metallic ores, unburnt carbon and ash particles are released. Research has indicated that the burning of bio-fuels, in the form of wood, agricultural waste, and usage of dried animal manure in cooking stoves, is the major source of black carbon emissions in India. These bio-fuel emissions contribute to 42% of the total emissions (Venkataraman, 2005).

### 2.9.2 Constituents of Clean Air

The 10 gases that make up clean are nitrogen, oxygen, argon carbon dioxide other gases like neon, Helium, Krypton, Methane (CH\textsubscript{4}), Hydrogen, and Xenon ozone and radon. As shown in the Table 2.1 nitrogen forms the major component of the clean air with 78.1%, followed by oxygen with 20.9% argon 0.9% carbon dioxide
is 0.35% and other gases form 0.065%. Apart from these gases sulphur dioxide at
0.0002 particles per million (ppm), oxides of Nitrogen are 0.001 ppm and carbon
monoxide particles at 0.001 ppm are also found in the air.

Table 2.1: Constituents of Clean Air

<table>
<thead>
<tr>
<th>Constituent</th>
<th>Proportion in Clean Air</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nitrogen</td>
<td>78.10%</td>
</tr>
<tr>
<td>Oxygen</td>
<td>20.90%</td>
</tr>
<tr>
<td>Argon</td>
<td>0.90%</td>
</tr>
<tr>
<td>Carbon dioxide</td>
<td>0.35%</td>
</tr>
<tr>
<td>Other gases</td>
<td>0.07%</td>
</tr>
<tr>
<td>Sulphur Dioxide</td>
<td>0.0002 ppm</td>
</tr>
<tr>
<td>Oxides of Nitrogen</td>
<td>0.001 ppm</td>
</tr>
<tr>
<td>Carbon Monoxide</td>
<td>0.001 ppm</td>
</tr>
</tbody>
</table>

A chemical species new to the composition of air leads to the pollution of air. Any increase in the level of the constituents of air and the chemical combination can cause pollution of air. When the temperature increase nitrogen combines with oxygen to form a class of toxic compounds called nitrogen oxides. The simplest one is Nitrogen oxide others being dinitrogen oxide, nitrogen dioxide, and dinitrogen dioxide. Nitrogen dioxide is a toxic compound often seen in polluted cities.

2.9.3 Constituents of Polluted Air

The primary and major air pollutants from various sources are sulphur dioxide (SO$_2$), oxides of Nitrogen (NO$_x$), Respirable Suspended Particulate Matter (RSP-M/PM10), and Carbon Monoxide (CO). All these pollutants are released as primary pollutants from various sources but atmospheric and meteorological processes may transform them to secondary pollutants like Ozone.

Sulphur dioxide (SO$_2$)

Sulphur dioxide or (SO$_2$), is from the family of sulphur oxide gases (SO$_x$). (SO$_x$) gases are formed when fuels that contain sulphur, namely coal and oil, are burned,
and also when metals are extracted from ore or gasoline is extracted from oil. SO₂ is the constituent of highest concern which is used as a meter for the group of gaseous sulphur oxides (SOₓ). SO₂ dissolves to form acid in water vapor, and also interacts with other gases and particles in the air to form sulphates and products that can be harmful to people and their environment. SO₂ causes a wide range of health and environmental issues because of its reaction with other materials in the air. People with asthma or lung disease, the elderly, and children are particularly sensitive groups. The influences of sulphur dioxide are respiratory effects from gaseous SO₂, respiratory effects from sulphate particles, visibility impairment, acid rain, etc. (EPA, 2010b, 2004).

Oxides of Nitrogen (NOₓ)

The gases found in the air in which nitric oxide (NO) and nitrogen dioxide (NO₂) are the dominant forms are represent by NOₓ. The oxides of Nitrogen (NOₓ) are mainly formed because of oxidation of atmospheric nitrogen during the combustion processes mainly and less by oxidation of organic nitrogen in fuels. These oxides are emitted due to combustion at high temperatures of fuels used in industrial activities, residential and commercial heating. The fires in forest are also a large natural source of (NOₓ) (source: CPCB 2012). NOₓ when reacts with moisture, ammonia, and other compounds generates nitric acid vapor. The small particles of nitric acid vapor pierce deep into lung tissue thereby damaging it and causing premature death. Inhaling such particles may lead to respiratory diseases like bronchitis, and also aggravate existing heart diseases. The NOₓ from combustion is formed in the stratosphere due to the photolysis of nitrous oxide that destroys the ozone in the stratosphere which absorbs the ultraviolet light causing harm to life on earth (EPA, 2010b, 2004).

Particulate Matter (PM)

The term particulate matter (PM) refers to a mixture of small (typically < 10 µm) particles both solids and may also include liquid droplets suspended in air (EPA, 2010b). Particulates may be emitted from natural and anthropogenic sources as primary air pollutants, or as secondary air pollutants in the atmosphere (EPA,
Particulate matter can either be primary or secondary. Primary particulate matter retains the same chemical form when released in the atmosphere. This includes dust such as road dust, fly ash, and soot blow due to wind. Secondary particulate matter is formed due to some chemical reactions in the atmosphere which includes nitrates and sulphates. Environmental effects of particulate matter include a reduction in visibility, aesthetic damage etc. Particulate matter can be categorized into two depending on the size of the particles which may be large with a diameter of 2.5 µm and 10 µm (PM10) and coarse fine particulates with diameters less than 2.5 µm (PM2.5). PM10 can be formed by physical processes like grinding, crushing and scraping and smoothening of surfaces. The larger size particles may be formed due to agricultural and mining activities. The effects of PM due to its exposure are largely experienced by people living in urban areas in rapidly developing and developed countries. The size of the particles is responsible for causing health problems. The particles that are 10 micrometers (µm) or smaller in diameter pass through the nose and the pharynx and go into the lungs. These particles once inhaled can cause severe health problems effecting the heart and lungs (USEPA, 2008).

When the size to these particles is less than 10 µm in diameter (PM10) they are called Respirable Suspended Particulate Matter (RSPM). The very fine particles have a higher propensity to remain air borne. These particles can be easily inhaled thus posing an excessive health hazard. These particles then collected and gathered in the air sacks of the lungs reducing the exchange of carbon dioxide and oxygen in the blood. The fine particles having a diameter less than 2.5 µ m get deep into the lungs and the bloodstream and thus are considered to be the greatest problems. Several studies and research have related particle pollution exposure to different health problems which may range from respiratory symptoms, problems faced during breathing, coughing, weakened lung function, exaggerated asthma, chronic bronchitis; irregular heartbeat and premature death in people with heart or lung disease (USEPA, 2008).

The various sources of RSPM are diesel generating sets, motor vehicles, power plants, fuel burning in industries, industrial furnaces and boilers,. Some other sources that may emit RSPM are solid waste disposal, construction activities, and traffic/road dust. Various data mining and interpolation techniques have help the
researchers to predict and estimate the values of air pollutants in the atmosphere based on the known values. Wonga et al (2004) selected four different interpolation methods to estimate O₃ and PM10 air concentrations. The four methods were spatial averaging, nearest neighbor, inverse distance weighting, and kriging (Jha et al., 2011).

2.10 Air Quality Index

The AQI is an index for reporting daily air quality. It tells you how clean or polluted your air is, and what associated health effects might be a concern for you. The AQI focuses on health affects you may experience within a few hours or days after breathing polluted air. AQI is value between 0 to 500. The higher the AQI value, the greater the level of air pollution and the greater the health concern. The AQI is an "index" determined by calculating the degree of pollution in the city or at the monitoring point and includes five main pollutants - particulate matter, ground-level ozone, sulfur dioxide, carbon monoxide and nitrogen dioxide. Each of these pollutants has an air quality standard which is used to calculate the overall AQI for the city. For better understanding and presentation, the AQI is broken down into six categories, each color coded with the number scale.( www.imd.gov.in/ ) The measurement scale is based on color system and a definite scale as per the Table 2.2. The AQI of each air pollutant is calculated using the Eq.2.3

\[
I_p = \left[ \frac{(I_{hi} - I_{low})}{BP_{hi} - BP_{low}} \right] \times (C_p - BP_{low}) + I_{low}
\]

(2.3)

where

\[ I_p \] is the index of the pollutant;

\[ C_p \] is the rounded concentration of pollutant p;

\[ BP_{hi} \] is the breakpoint greater or equal to \[ C_p \];

\[ BP_{low} \] is the breakpoint less than or equal to \[ C_p \];
\( I_{hi} \) is the AQI corresponding to \( BP_{hi} \);

\( I_{low} \) is the AQI corresponding to \( BP_{low} \);

### Table 2.2: Air Quality Index

<table>
<thead>
<tr>
<th>AQI</th>
<th>Color</th>
<th>Air Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-50</td>
<td>Green</td>
<td>Air quality is considered satisfactory and air pollution poses little or no risk</td>
</tr>
<tr>
<td>51-100</td>
<td>Yellow</td>
<td>Air quality is acceptable with some pollutants, which might cause some problem to health</td>
</tr>
<tr>
<td>101-150</td>
<td>Orange</td>
<td>Air quality is unhealthy for some people known as the sensitive group. The general people will not be affected</td>
</tr>
<tr>
<td>151-200</td>
<td>Red</td>
<td>Quality of air is unhealthy for all, causing a health hazard for the masses in general</td>
</tr>
<tr>
<td>201-300</td>
<td>Brown</td>
<td>Air quality is very unhealthy and is in a state of emergency</td>
</tr>
<tr>
<td>&gt;300</td>
<td>Black/ Grey</td>
<td>The condition of the air is hazardous with a general health hazard experienced by all</td>
</tr>
</tbody>
</table>

### 2.11 Air Pollution and Meteorology

Meteorology plays a vital role in influencing the concentrations of air pollutants in the atmosphere at a given location. The vital way in which it can effect is the dispersion or the dilution of the pollutants. It governs the dispersal completely once the pollutants are emitted into the atmosphere. (www.imd.gov.in/) Studies have shown that many of the air pollution episodes are have occurred under hostile meteorological conditions. An effective air pollution warning can be through
a meteorological forecast from which a mitigation strategy can be worked out. Cautions and alerts based on air quality alone will be useless if the variable atmospheric processes of pollutants dispersal are not taken into account. The wind is an important meteorological variable that can be used in air pollution studies. The direction and the speed of transport of air pollutants are governed by the wind. The wind speed determines to some extent the dilution of the pollutants in the atmosphere, the direction gives the carriage of the air pollutants in the atmosphere. An equally important role is played by the temperature variation in air pollution. The increase of temperature at a low level prevents the dispersion of the air pollutants and results in enormously high ground level concentrations. Studies have indicated that there is strong correlation between the air pollutants and the meteorological elements. Air pollution episodes are caused due to various factors which include emissions, local and synoptic scale and location, meteorological conditions, topography, and atmospheric chemical processes. As per the studies the strongest correlations occur between NOx with temperatures and pressure. Higher NOx occurs when temperatures are lower and pressures are higher. Increasing wind speed helps air pollutants dispersion and decreases its level (Gardner and Darling, 1996, 1999 and 2000 ; Erika K., 2005).

2.12 Air Pollution Prediction and Forecasting

Various data mining techniques have been used by various researchers in the air pollution to study the behavior of air pollutants in the atmosphere. Sheng-Tun Li et al (2006) has assessed one of the spatiotemporal data mining techniques, the cluster analysis, on air pollution. Multi-scale input data was used as an input to the SOM neural networks and the spatio-temporal change was studied. The study showed that the number of clusters reduced as the scale of data input increases. This indicates that the scale is an important factor in the cluster analysis. The study showed that Self Organized Maps (SOM) is an excellent visual tool for studying the inner structure of the data transformed and provides the capability for hard clustering. The various advantages and limitations of SOM, fuzzy SOM, fuzzy K-means, were studied. The fuzzy logic-based methods use the membership to quantitatively measure the association between one station and each cluster. The
threshold assigned to the membership provides the capability to handle clustering when it is difficult to assign one station to a specific cluster since it could have close relationship with several adjacent clusters.

Vincent et al. (2001) implemented spatial data mining of air pollution data and proposed an agent framework to study the consequence of meteorological elements on air pollutant elements. The framework consisted of three components which included agents to collect the data, coordinate and querying. The user interacts with the Querying Agents and sends queries and mining request which then deliver the results. Coordinating agents get the data resources ready for the querying task and apply spatial data mining methods. The knowledge discovery proposes involves data mining process where the knowledge exists in the form of concept hierarchies. The proposed method uses the concept hierarchies explicitly known by the experts, or generated automatically by data study.

Nikov et al. (2005) proposed AirPolTool; a web-based tool for predicting air pollution in Istanbul. The tool models the association between local meteorological data and concentrations of air pollution indicators like SO$_2$, PM10 and CO with the help of neural networks. The tool is very easy to use giving a three-days prediction of air pollutant and necessary warning signals and appropriate actions to be taken by the administration to reduce the level of the particular air pollutant to a level so that it is no longer harmful.

Deleawe et al. (2011) explore the use of machine learning technologies to forecast CO$_2$ levels as a measure for the quality of air. The machine learning technologies are used to examining the correlation between resident activities and air quality levels. The methods presented study the sensor information in smart environments. The efficiency of these methods is assessed using three smart environment test beds. The first technique used is a nave Bayes classifier which gave promising results for recognition of a particular activity. The second model used a multilayer perceptron that accepts continuous-valued attributes like temperature. The third model used the decision tree to select an ordering of feature values. The models developed are implemented using the open source software for data mining Weka. This study proved the ability of machine learning algorithms to forecast air quality levels in smart environment sensor information. The results obtained would further be used to develop measures to reduce the affect and purify the air.
Athanasiadis et al. (2007) used classification techniques, in an environmental management system to help the decision making. The system supervises the ambient air quality and generates warnings in case of occurrence of events. They used an extensive data set from the real world which has ensured reliable prediction procedures that are capable of taking decisions at the operational level. These methods have added value to the traditional approaches in air quality assessment, thus improving the performance of the forecasting methods. A C4.5 algorithm for decision tree generation was applied in WEKA and the study produced reliable models that had a forecasting accuracy exceeding 93%. This classification approach made it possible to handle data uncertainties involved in an air quality and thus support decision makings in an operating timeframe.

Keith McCabe et al. (2007) made use of statistical methods to analyze local air quality problems in Europe. The modern statistical technique Generalized Additive Models (GAMs) was used on detailed traffic information to model and forecast concentrations of air pollutants near the motorways. The models developed were also then employed to check for trends in concentrations keeping the climatic influence of meteorology conditions constant. The GAM models were further used to study the importance of Primary NO\textsubscript{2} at the specific sites and were compared with the results of a chemical model that forecasted trend in primary NO\textsubscript{2} from vehicle emissions on the motorway. The comparison identified the primary as well as the secondary causes of pollution in the area. The mitigation measures focused on the primary sources and then on the secondary sources at a very early stage of the project since they used the data mining process. This project helped the decision makers justify the development of mitigation measures at an earlier stage in the project programme. This proved that there is a possibility to use these techniques and link the multi-source data collection process as a tool to support decision making and help the decision makers to study the impacts of previous strategies.

Ma et al. (2010) and Guo et al. (2008) analyzed air pollution pattern in London based on K-means algorithm as distributed and centralized techniques. The results were compared and the distributed approached proved to be more effective than the centralized technique. This research also explores the dispersion of different pollution clouds to see their movements and changes. The pollutants of NO\textsubscript{x}...
(NO+NO₂) and SO₂ respectively were analyzed and the pollution clouds were estimated at three different time intervals with a gap of 15 minutes.

Yajie Ma et al. implemented a distributed system using data mining, grid computing and wireless sensors to develop a cost effective approach. The model developed was used to collect real time environmental data and check air pollution in urban regions. The aim was to model and forecast air pollutants and dangers using a grid of sensors fixed along the roadsides. The system provides a model for air pollution in urban environment giving a round the clock information of the air pollution disparity and environmental protection.

U. W. Tang estimated the air quality for each individual building along the sides of the road and compared it with the grid-based approach. The approach forecasted the compound spatial variation of urban geometry, emission due to traffic, air pollution and its dispersion. The aim was to model and predict a range of air and environmental pollutants and dangers using a grid of sensors. The system offers an air pollution pattern in urban environment enabling real-time tracking of variation in the air pollution. This study shows that the approach may lead to a methodology in data mining of urban environment.

Chen et al. (2007) put forward a Decision-Making Framework (DMF) for studying the ozone pollution in the metropolitan region of Atlanta. High concentrations of ozone are a problem in numerous US cities, Atlanta being among the top cities on the list. The air chemistry of ozone was developed using a mining and metamodeling tool that was computationally efficient. The proposed approach effectively modeled the changes in ozone concentrations over a period of one day. The Atmospheric Chemistry Module for ozone pollution DMF implemented successfully for ozone monitoring in Atlanta.

Yu-Kai Huang et al. (2013) studied use of multiple regression analysis on the spatial distribution of SO₂ and NO₂ pollution in Ulaanbaatar, Mongolia. This study draws the conclusion that in Ulaanbaatar the ambient air pollution was high during the heating season which was associated with pollution levels of SO₂ and NO₂.

Nastos et al (2013) estimated the probability of predicting the daily precipitation for the following year. Artificial Neural Networks (ANNs) was used to assess the daily extreme precipitation forecast. The techniques of coefficient of deter-
mination ($R^2$), the Root Mean Square Error (RMSE) and the Mean Bias Error (MBE), were used to test the reliability of the models. They proved that the proposed ANN model worked very well with the historical data.

Fernando et al. (2002) used stochastic methods to transfer the decision making to ANN and predict the particulate matter concentration in Phoenix, Arizona and found that it better predicts PM10 concentrations. ANN can be very useful in developing rapid air quality warning systems and they neither require expensive inventory nor regular upgrading.

P. Viotti et al. (2002) forecasted short and middle term concentration levels for air pollutants using the ANN. The ANN showed good performances using the back propagation algorithm. They developed a model on the basis of a hypothesis about the values of the meteorological parameters and tried it out on the city of Perugia.

Martin et al. (2008) employed artificial neural networks (ANNs) using the Multilayer perceptron models (MPLs) with back propagation learning rule and $k$-nearest Neighbors ($k$-nn) classifiers for predicting high levels of ambient carbon monoxide. Errors in prediction were found to be 20% for the models based on ANNs. Jeong-Sook Heo et al. (2004) developed two types of forecasting models based on fuzzy logic and neural networks for the city of Seoul, South Korea for forecasting the high ozone levels occurring on the next day. They suggested that the precision of the predicting system has been upgraded continuously through authentication and increase in the availability of the data.

Gokhale et al. (2008) developed three air quality models modified General Finite Line Source Model (M-GFLSM) of particulates, the California Line Source (CALINE3) model, and the California Line Source for Queuing Hot Spot Calculations (CAL3QHC) to evaluate the daily average PM10 and PM2.5 concentrations.

Dimitris Voukantsis et al. (2007) propose a methodology to compare the meteorological data and air quality for predicting the air pollutants of interest in the urban areas based on computational intelligence methods, principal component analysis and artificial neural networks. They formulated a hybrid scheme of linear regression and ANN models for developing air quality forecasting models.

Gulliver et al. (2011) proposed an air pollution model to forecast annual and
daily levels of PM10 based on the geographic Information System. The model used a grid-based function implemented in ArcGIS using the data on meteorology to implement a simple Gaussian plume model of air pollution dispersion. The proposed approach achieved value of R2 between 0.40 to 0.77 for annual daily predictions. This model was useful for speedy production of day-to-day or yearly city-wide air pollution maps that helped in urban air quality forecasting, planning and management.

Davor Z. Antanasijevi et al. (2013) developed an artificial neural network (ANN) model for the prediction of yearly PM10 emissions. The ANN model has performed exceedingly well and has proven that the forecast of PM10 emission for the next two years can be made effectively and precisely. The mean absolute error for two-year PM10 emission prediction was only 10%. This is three times better than the forecasts obtained from the conventional principal component regression models and multi-linear regression that were trained and tested using the same datasets and input variables.

Kurt et al. (2010) used geographic forecasting models using neural networks (GFM_NN) for forecasting particulate matter (PM10) carbon monoxide (CO), and sulfur dioxide (SO2) that are responsible for urban air pollution. Results have shown very less error for real time forecasting system and are easily adaptable.

Domanska D. et.al. (2012) proposed a model based on fuzzy logic and expert systems to forecast the concentrations of PM10, PM2.5, SO2, NO, CO and O3 for a selected number of hours later using meteorological data. The errors were below 30% and proved to be satisfactory in the areas of Poland.

Wise et al. (2005) concludes which meteorological variables most effect ozone and PM in the Southwest and inspects patterns of underlying pollutant trends due to emissions. According to their study among all the meteorological variables namely the temperature, humidity, rainfall pressure, wind speed temperature plays the most significant role in the concentrations of ozone. The relative humidity is the strongest predictors of the air pollutants PM.

Chelani et al. (2006) developed a hybrid methodology that can deal with both the linear and nonlinear structure of the air pollutants. They used the autoregressive integrated moving average model to deal with the nonlinearity and the dynamic nature. It was observed that the performance of the hybrid model was
better than the linear and nonlinear individually thus making it a powerful tech-
nique to predict the air pollutants concentrations.

Fasbender et al. (2009) used the Bayesian data fusion (BDF) framework to predict the air pollutants using the several secondary sources of information. The framework used lead to a greater understanding of the information and behavior of the air pollutant and gives better predictions of the air pollutant NO$_2$.

Corani (2005) used the Feed-forward neural networks (FFNNs), for forecasting air quality statistically and for forecasting ozone and PM10 in the city of Milan. They results of the FFNS were compared with the lazy learning and pruned neural networks. The performances of the lazy learning performed better on an average and pruned neural networks proved to the best for detection of variances from the threshold.

Nagendra et al. (2004) used the ANN technique to model the dispersion of nitrogen dioxide (NO$_2$) and showed an acceptable performance of the models based on ANN studied on the evaluation data set in Delhi city.

Jianga et al. (2004) developed an artificial neural network (ANN) model with meteorological forecasting data as the main input for API forecasting in Shanghai, to forecast the average API values for the next day.

Hooyberghs et al. (2005) developed a neural network tool to predict the average PM10 concentrations daily in Belgium one day in advance. They concluded that daily fluctuations of PM10 concentrations in urban areas of Belgian are mainly driven by meteorological conditions not by changes in anthropogenic sources.

Perez-Roa et al. used the artificial neural networks (ANN) and assessed how ANN model can improve the ability of dispersion to forecast highest concentrations of ambient carbon monoxide in a large city using meteorological data of 8 cities. Results showed that the ANN model attained better forecasts of highest CO concentration levels than before. The errors also reduced by half than before.

Brunellia et al. (2007) developed a PM2.5 air quality prediction model that is based on nonlinear regression (NLR) and back-trajectory concentrations. The models implemented back trajectory concentrations and compared to the other models showed lower mean absolute errors and high rates of detecting unhealthy PM2.5 concentrations.

Perez (2006) developed a PM10 model using ANN for forecasting and used it
for air quality management in Santiago, Chile. Air pollutant concentrations taken from several locations from the city and meteorological information in the region act as inputs to the models. Results obtained for the last three years are indicative of the fact that the model may be considered as an important tool for air pollution control.

Gautam et al. (2008) used the artificial neural network technique to give reliable prediction results for the city of Delhi. The techniques proposed to use the nearest neighbor technique for prediction from the training set. The results obtained were compared with the back propagation algorithm. The proposed scheme offers the capability to capture the dynamics that underlie the chaotic time series, since the input patterns are presented one after another to the network.

Pfeiffer et al. (2009) put forth a new method using artificial neural networks for calculating the average spatial distribution of air pollutants based on diffusive sampling measurements. They proved that interpolation algorithms do not adapt themselves to the changes in the dispersal parameters. A multilayer perceptron was trained with the NOx measurement data and distribution parameters like meteorological parameters were considered. The best fit was attained with meteorological parameters as input to the nodes of the neural network, resulting in accurate maps of the annual average distribution of NO2.

Central Pollution Control Board (CPCB) has pointed out in one of its reports Mumbai is one of the cities where the main pollutants reported are RSPM and Suspended Particulate Matter. Mumbai being a coastal city and most of its population living in the low lying areas, people will be greatly affected by climate change (MCGM, 2011). According to latest studies undertaken by the Maharashtra Pollution Control Board (MPCB) in 2010, air pollution levels in the city are approaching the point of no return. NEERI (2010) has also pointed out that it is necessary to understand that high load contribution does not necessarily contribute to the high ambient contribution of the particular parameter at the receptor site. This is due to the fact that emission distribution in atmosphere is not dependent on a single factor but multiple factors such as location, meteorological conditions, atmospheric removal processes and daily variation. Various studies carried out in the past decade show that Mumbai is probably the highest vulnerable city in terms of metrological variables and climatic changes. A forecasting system which
provides information on Air Quality with respect to specific location in real time is needed. As an initiative the Indian Institute of Tropical Meteorology (IITM), Pune, a principal under the Ministry of Earth Sciences, Government of India, has come up with the country’s first major project named as ”System of Air quality Forecasting And Research (SAFAR)”. The SAFAR system provides location specific information on Air Quality in real time and predicts the concentrations of air pollutants one day in advance. It is supported by the weather forecasting system designed by IMD, New Delhi. The ultimate objective is to increase awareness among the general public regarding the air quality in respective cities well in advance so that appropriate mitigation action and systematic measures can be taken up for the improvement of air quality and related health problems. SAFAR system assimilates various complex components like meso-net monitoring network consisting of online air pollution analyzers, automatic weather stations, emission inventory, activity data; 3-D coupled atmospheric chemistry transport models to facilitate prediction of several major air pollutants like: ozone, NOx, CO, PM2.5, PM10, benzene, toluene, xylene and BC. This model has been successfully tested and executed during the Commonwealth Games held at New Delhi in 2010. The same approach is to be followed in the other 5 major cities of India namely Ahmedabad, Kolkata, Mumbai, Pune, Chennai and Bengaluru. [source: www. (http://safar.tropmet.res.in/)]

2.13 Summary of Literature Survey Done

The Table 2.3 summarizes the literature survey done along with the study and research done and the various techniques, algorithms and the findings of the various researchers internationally and in India.
<table>
<thead>
<tr>
<th>No</th>
<th>Author</th>
<th>Year</th>
<th>Discussion</th>
<th>Techniques/Algorithms/Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Nastos et al.</td>
<td>2013</td>
<td>Maximum daily precipitation for the next year</td>
<td>Artificial Neural Networks to evaluate the possibility of daily extreme precipitation forecast; proved that the proposed ANN model worked very well with the historical data.</td>
</tr>
<tr>
<td>2</td>
<td>Athanasiadis et. al.</td>
<td>2013</td>
<td>Supported the decision making process involved in environmental management in order to monitor the ambient air quality and trigger warnings in case of emergency</td>
<td>Classification technique &amp; Generalized Additive Models</td>
</tr>
<tr>
<td>3</td>
<td>Yu-Kai Huang</td>
<td>2013</td>
<td>Study spatial distribution and source contribution of SO₂ and NO₂ pollution in Ulaanbaatar, Mongolia</td>
<td>Multiple regression models</td>
</tr>
<tr>
<td>4</td>
<td>Davor Z. Antanasijevi et al.</td>
<td>2013</td>
<td>Forecasting of annual PM10 emissions at the national level</td>
<td>Artificial neural network (ANN) model</td>
</tr>
<tr>
<td>5</td>
<td>Fernando et al.</td>
<td>2012</td>
<td>Predict the concentration of particulate matter at a monitoring site in Phoenix, Arizona at regular intervals; better predicted the PM 10 concentrations than continuous monitoring</td>
<td>ANN with back propagation</td>
</tr>
</tbody>
</table>
Table 2.3 – Continued from Previous Page

<table>
<thead>
<tr>
<th>No</th>
<th>Author</th>
<th>Year</th>
<th>Discussion</th>
<th>Techniques/Algorithms/Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>D. Domanska et al.</td>
<td>2012</td>
<td>Predict the concentrations of particulate matter PM10, PM2.5, SO₂, NO, CO and O₃ for a selected number of hours ahead using meteorological data and the errors were below 30%</td>
<td>Fuzzy logic and expert systems</td>
</tr>
<tr>
<td>7</td>
<td>U. W. Tang</td>
<td>2011</td>
<td>Estimated air quality in front of building along both sides of the road</td>
<td>The building based approach in order to increase the spatial resolution of the data</td>
</tr>
<tr>
<td>8</td>
<td>Dimitris Voukantsis et al.</td>
<td>2011</td>
<td>Compared the air quality and meteorological data to forecast the air pollutants of interest in Thessaloniki and Helsinki in Greece and Finland. They formulated a hybrid scheme of linear regression and ANN models for development of air quality forecasting models.</td>
<td>principal component analysis, artificial neural networks and computational intelligence methods,</td>
</tr>
<tr>
<td>9</td>
<td>John Gulliver</td>
<td>2011</td>
<td>Forecast the annual and daily levels of PM10</td>
<td>GIS-based air pollution model using grid-based function in ArcGIS</td>
</tr>
<tr>
<td>No</td>
<td>Author</td>
<td>Year</td>
<td>Discussion</td>
<td>Techniques/Algorithms/Findings</td>
</tr>
<tr>
<td>----</td>
<td>----------------</td>
<td>------</td>
<td>----------------------------------------------------------------------------</td>
<td>------------------------------------------------</td>
</tr>
<tr>
<td>10</td>
<td>Seun Deleawe</td>
<td>2010</td>
<td>Analyze the temporal patterns and recognizing interleaved activities.</td>
<td>Hidden Markov model</td>
</tr>
<tr>
<td>11</td>
<td>Yajia Guo et al</td>
<td>2010</td>
<td>Measures, models and predicts environmental pollutants and hazards and provides a typical air pollution pattern in urban environment which gives a real-time track of the air pollution variation sensors.</td>
<td>Wireless sensors network and Grid computing technology</td>
</tr>
<tr>
<td>12</td>
<td>D. Fasbender et al.</td>
<td>2009</td>
<td>Better understanding of the information and behavior of the air pollutant and gives better predictions of the air pollutant NO₂.</td>
<td>The Bayesian data fusion (BDF) framework</td>
</tr>
<tr>
<td>13</td>
<td>H. Pfeiffer et al</td>
<td>2009</td>
<td>The best fit was achieved with meteorological parameters as input nodes for the neural network, predicts the annual average distribution of NO₂ in Cyprus.</td>
<td>On diffusive sampling measurements and artificial neural networks</td>
</tr>
<tr>
<td>14</td>
<td>Martin et al.</td>
<td>2008</td>
<td>Predictive tool for high levels of ambient carbon monoxide (CO).</td>
<td>Artificial Neural Networks</td>
</tr>
<tr>
<td>No</td>
<td>Author</td>
<td>Year</td>
<td>Discussion</td>
<td>Techniques/ Algorithms/ Findings</td>
</tr>
<tr>
<td>----</td>
<td>--------</td>
<td>------</td>
<td>------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td>15</td>
<td>Keith McCabe</td>
<td>2007</td>
<td>Able to identify the key elements of pollution in the area and estimate the importance of the various sources; helps the decision makers by giving statistical data to develop mitigation measures</td>
<td>Generalized Additive Models (GAMs) coupled with traffic information to model and predict air pollutant concentrations near the roads</td>
</tr>
<tr>
<td>16</td>
<td>Sharad Gokhale et al.</td>
<td>2008</td>
<td>Evaluate the daily average PM10 and PM2.5 concentrations.</td>
<td>Developed modified General Finite Line Source Model (M-GFLSM) of particulates, the California Line Source (CALINE3) model, and the California Line Source for Queuing &amp; Hot Spot Calculations (CAL3QHC)</td>
</tr>
<tr>
<td>17</td>
<td>Ajit Kumar Gautam et al.</td>
<td>2008</td>
<td>Artificial neural network technique to give reliable prediction results for the city of Delhi</td>
<td>The network captures the dynamics of the uneven time series, as the input patterns are given one by one to the network.</td>
</tr>
<tr>
<td>18</td>
<td>Chen et.al.</td>
<td>2007</td>
<td>Reducing ozone pollution in the metropolitan areas</td>
<td>Mining and meta modeling tools</td>
</tr>
<tr>
<td>No</td>
<td>Author</td>
<td>Year</td>
<td>Discussion</td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>------------------------</td>
<td>------</td>
<td>-----------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>A. C. P. Chen</td>
<td>2007</td>
<td>Decision-Making Framework (DMF) for reducing ozone pollution in the metropolitan Atlanta region</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mining and metamodeling tools and presented the results effectively to model the changes in the ozone concentrations over a period of 24 hours</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>U. Brunellia et al</td>
<td>2007</td>
<td>Developed a PM2.5 air quality forecast model</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Back-trajectory concentrations and Nonlinear regression (NLR); model had lower mean absolute errors and better rates of detecting unhealthy PM2.5 concentrations compared to the other models.</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Wang et al.</td>
<td>2006</td>
<td>Forecasted the Air Pollution Index (API)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Autoregressive moving average (ARMA)</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Sheng-Tun Li et al.</td>
<td>2006</td>
<td>Membership or correlation coefficients provides the capability to handle clustering when it is difficult to assign one station to a specific cluster since it could have close relationship with several adjacent clusters.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Spatiotemporal data mining techniques ; cluster analysis</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>Author</td>
<td>Year</td>
<td>Discussion</td>
<td>Techniques/ Algorithms/ Findings</td>
</tr>
<tr>
<td>----</td>
<td>-----------------------</td>
<td>------</td>
<td>----------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>23</td>
<td>Asha B. Chelani et al.</td>
<td>2006</td>
<td>Hybrid methodology that can deal with both the linear and nonlinear structure of the air pollutants.</td>
<td>Autoregressive integrated moving average model to deal with the nonlinearity and the dynamic nature &amp; hybrid model performed better than the linear and nonlinear individually thus making it a powerful technique to predict the air pollutants concentrations.</td>
</tr>
<tr>
<td>24</td>
<td>Patricio Perez</td>
<td>2006</td>
<td>Developed a PM10 forecasting model for air quality management using ANN in Santiago, Chile.</td>
<td>Results for three years indicate that the model is an important tool for air pollution control.</td>
</tr>
<tr>
<td>25</td>
<td>Heo</td>
<td>2005</td>
<td>Forecast air pollutants based on meteorological variables. The neural network model was improved continuously through verification and augmentation.</td>
<td>Fuzzy expert and neural network systems</td>
</tr>
<tr>
<td>26</td>
<td>Nikov et. al.</td>
<td>2005</td>
<td>Relates the local meteorological data and air pollution indicators concentrations like SO₂, PM10 and CO</td>
<td>ANN ; a web-based tool for air pollution prediction</td>
</tr>
<tr>
<td>No</td>
<td>Author</td>
<td>Year</td>
<td>Discussion</td>
<td>Techniques/ Algorithms/ Findings</td>
</tr>
<tr>
<td>----</td>
<td>----------------------------</td>
<td>------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------</td>
</tr>
<tr>
<td>27</td>
<td>Atakan Kurt et al.</td>
<td>2005</td>
<td>Forecasting carbon monoxide (CO), sulfur dioxide (SO₂) and particulate matter (PM10) that is primarily responsible for air pollution in urban areas that is difficult to forecast</td>
<td>Geographic forecasting models using neural networks</td>
</tr>
<tr>
<td>28</td>
<td>Erika K. Wise</td>
<td>2005</td>
<td>Examines patterns of underlying pollutant trends due to emissions</td>
<td>Meteorological variables most influence ozone and PM in the Southwest</td>
</tr>
<tr>
<td>29</td>
<td>Giorgio Corani</td>
<td>2005</td>
<td>Statistical prediction of air quality prediction of PM10 and ozone in Milan</td>
<td>Feed-forward neural networks (FFNNs),</td>
</tr>
<tr>
<td>30</td>
<td>Jef Hooy - berghs et al.</td>
<td>2005</td>
<td>To forecast a day in advance the daily average PM10 concentrations in Belgium.</td>
<td>Neural network tool to predict PM10 concentrations in Belgian urban areas; are dependent on meteorological conditions and not anthropogenic sources</td>
</tr>
<tr>
<td>31</td>
<td>S.M. Shiva Nagendra et al.</td>
<td>2004</td>
<td>Modeled the nitrogen dioxide (NO₂) dispersion phenomena</td>
<td>Artificial neural network (ANN) model</td>
</tr>
<tr>
<td>32</td>
<td>Viotti et al.</td>
<td>2002</td>
<td>Forecast short and middle term concentration levels for some of the well-known pollutants</td>
<td>ANN &amp; perceptron with backpropagation algorithm model</td>
</tr>
</tbody>
</table>
Table 2.3 – Continued from Previous Page

<table>
<thead>
<tr>
<th>No</th>
<th>Author</th>
<th>Year</th>
<th>Discussion</th>
<th>Techniques/ Algorithms/ Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>33</td>
<td>Vincent Ng. et al.</td>
<td>2001</td>
<td>Agent framework to study the effect of meteorological data on the air pollutants</td>
<td>Statistical Data Mining</td>
</tr>
<tr>
<td>34</td>
<td>Dahe Jia et al.</td>
<td>2001</td>
<td>The API forecasting the next day average API values in Shanghai with meteorological data as the main input</td>
<td>Artificial Neural Networks</td>
</tr>
<tr>
<td>35</td>
<td>R. Perez-Roa et al.</td>
<td>2001</td>
<td>ANN model improves the capability of dispersion to predict peak concentrations of ambient carbon monoxide in a large city using meteorological data of 8 cities</td>
<td>ANN model achieved better predictions of peak CO concentration levels than before. The errors also reduced by half than before.</td>
</tr>
<tr>
<td>36</td>
<td>Central Pollution Control Board (CPCB)</td>
<td></td>
<td>Reports and Manuscripts</td>
<td>Standards set for the air pollutants</td>
</tr>
</tbody>
</table>

### 2.14 Tools for Implementation

Neural networks can be simulated using software applications dealing with a limited number of neural networks. The practical applications artificial neural networks used for data analysis are the neural networks that emphasis on data mining and forecasting. These simulators have the capability of preprocessing to develop a static neural network that can further be altered and configures according to the requirements. The data analysis simulators normally use self-organizing maps or the back propagating networks as their basic type of artificial neural networks.
Matlab's Neural Network toolbox supports supervised and unsupervised networks. The supervised networks consist of the multilayer feed forward, radial basis, learning vector quantization (LVQ), time-delay, nonlinear autoregressive (NARX), and layer-recurrent types of networks. The unsupervised networks include competitive layers and the self-organizing maps. The training and learning functions adjust the network's weights and biases automatically based on mathematical functions. Neural Network Toolbox supports numerous training algorithms that include gradient methods, the Levenberg-Marquardt algorithm (LM), and the back propagation algorithm. Customized custom training algorithms can be developed using the toolboxes integrated framework and can be integrated with built-in algorithms.[source: www.mathworks.in/help/]

Neuroph is lightweight Java neural network framework that can be used to develop neural network architectures with the coding language as Java. It contains, open source Java library with limited number of basic classes required for the neural networks.

Weka is a set of machine learning algorithms for accomplishing various data mining tasks. The algorithms can directly implement on a dataset or a Java code could be written to call the data set. Weka has a set of tools for various techniques for data pre-processing, clustering, classification, association rules, regression, and visualization. The software is well-suited for developing new machine learning systems and is designed for data processing where training data can be loaded into the memory and processed.

Fast Artificial Neural Network Library is a free open source neural network library that implements multilayer artificial neural networks using the C language and supports fully and sparsely connected networks. It also provides a method for easy handling of training data sets.

NearPy is a simple yet modular little framework for ANN search written in the very popular programming language Python.

Kriging can be implemented using various tools like are R packages, ArcGIS, Matlab, SCilab and Python. The kriging tool offered in Mathlab does not provide a library which can user created programs. It proves to be a good solution for small number of data set and not suitable for automating the kriging process for a large data set. The R package geoR uses the R software to provides functions
for geostatistical data analysis. It provides features of custom made functions to produce empirical variogram estimates that take the distance and the direction into consideration for a geostatistical data analysis.

2.15 Conclusions

After an exhaustive review and survey of literature related to the spatial data mining techniques, kriging, effects of air pollutants, effects of metrological data on air pollution, it was found that various data mining techniques namely cluster analysis, regression analysis, neural networks, fuzzy logic have been used in major cities worldwide for forecasting of the air pollutants. It was seen that not much attention is paid to hybrid models or combining two or more techniques to further improve the data mining results. The literature provides many relevant insights to the use of artificial intelligence techniques namely Artificial Neural Networks because of its main advantage of dealing with nonlinear problems. The literature review examined the published data on various interpolation techniques that can be used to forecast the values of air pollutants at unmonitored cites. It was concluded that the kriging is by far a good interpolation technique in location based data because it considers the spatial heterogeneity and autocorrelation into consideration while estimating and forecasting the parameters at unmonitored locations.

Most of the literature review found focused on using a single technique for forecasting air pollution for eg. Artificial neural networks to forecast the parameter in a particular domain. Very few researchers have made an attempt to combine two or more techniques and have come up with promising results. The applications of ANN related to the air pollutants and its forecasting was directed towards forecasting one of the air pollutants. Very few instances of developing an ANN for multiple pollutants were seen.

There is lot of literature was reviewed to study the sources and reasons for the increase in air pollutants in many urban areas around the globe. The literature on the effects of meteorological parameters on the air pollutants was examined and it was found that meteorological parameters do play an important role on the dispersion of the air pollutants namely temperature, humidity and wind speed.
Also Urban Air pollution Management Systems are the need of the day. The literature review examined the extensive published data to monitor the air quality in cities well in advance so that appropriate mitigation action so that precise measures can be taken up for the related health issues. It was concluded that Mumbai is likely to be highly vulnerable to climate change and meteorological variables. Although there is already lots of literature, there is some study required in enhancing the knowledge of combining different techniques to come up with a hybrid technique in order to predict and forecast the air pollutants in the area of Mumbai and Navi Mumbai effectively. There have been limited cases of forecasting and prediction using hybrid techniques in Mumbai. Also there is abundance of location based i.e. spatial information available with the MET department as well as the Maharashtra Pollution Control Board which not shared amongst each other for forecasting.

It was found that negligible work is done for air pollutants forecasting using data mining techniques and around the locations of Mumbai and Navi Mumbai considering the meteorological parameters. Also forecasting of the air pollutants apart the areas of the monitoring stations are not estimated and considered. This motivated the researcher to correlate the meteorological data of Mumbai from the MET Department and air pollutants namely SO$_2$, NO$_x$ and RSPM by obtaining the data from MPCB to integrate, predict and forecast the air pollutants in different parts of Mumbai and Navi Mumbai using the hybrid combination of Artificial Neural Networks and Kriging.